Regression Analysis of Car Price Prediction

Presented by-
Sanskruti Sonawane - 23060641041
Arpan Dey - 23060641013

Report
(Linear Models)

We have chosen a data-set which is an previous sales of the used Toyota Corolla cars at the dealership. The data set contains 1459 rows with 39 columns. The columns consists of 35 integer column, 3 String columns and 1 ID column. The columns represents the variables which are namely Id, Model, Price, Age, Mfg_Month, Mfg_Year, KM, Fuel_Type, HP, Met_color, Color, Automatic, CC, Doors, Cylinders, Gears, Quaterly Tax, Weight, Mfr_Guarantee, BOVAG_Gurantee, ABS, Airbag 1, Airbag 2, Airco, Automatic Airco, Boardcomputer, CD Player, Central Locks, Powered windows, Power steering, Radio, Mistlamps, Sport Model, Backseat divider, Metallic Rim, Radio Cassette, Parking Assistant, Tow Bar. Here, we observe that Price of the car is completely dependent on other independent variables Age_08_04, Mfg_Year, KM, Weight, Automatic_airco, Boardcomputer.

So, We on next we frame the problem statement which is as follows:-

- ## Problem Statement :

To build a regression model to predict **the Price of Car** with respect to the most contributing 6 independent variables i.e. **'Age_08_04', 'Mfg_Year', 'KM', 'Weight', 'Automatic_airco', 'Boardcomputer'** by examining it.

- ## Dataset:

the embedded data file is attached below:-

https://www.kaggle.com/datasets/victorahaji/toyota-corolla-car-price-prediction

- ## Hypothesis :

For the **regression model,** the hypothesis for simple linear regression is as follows :-

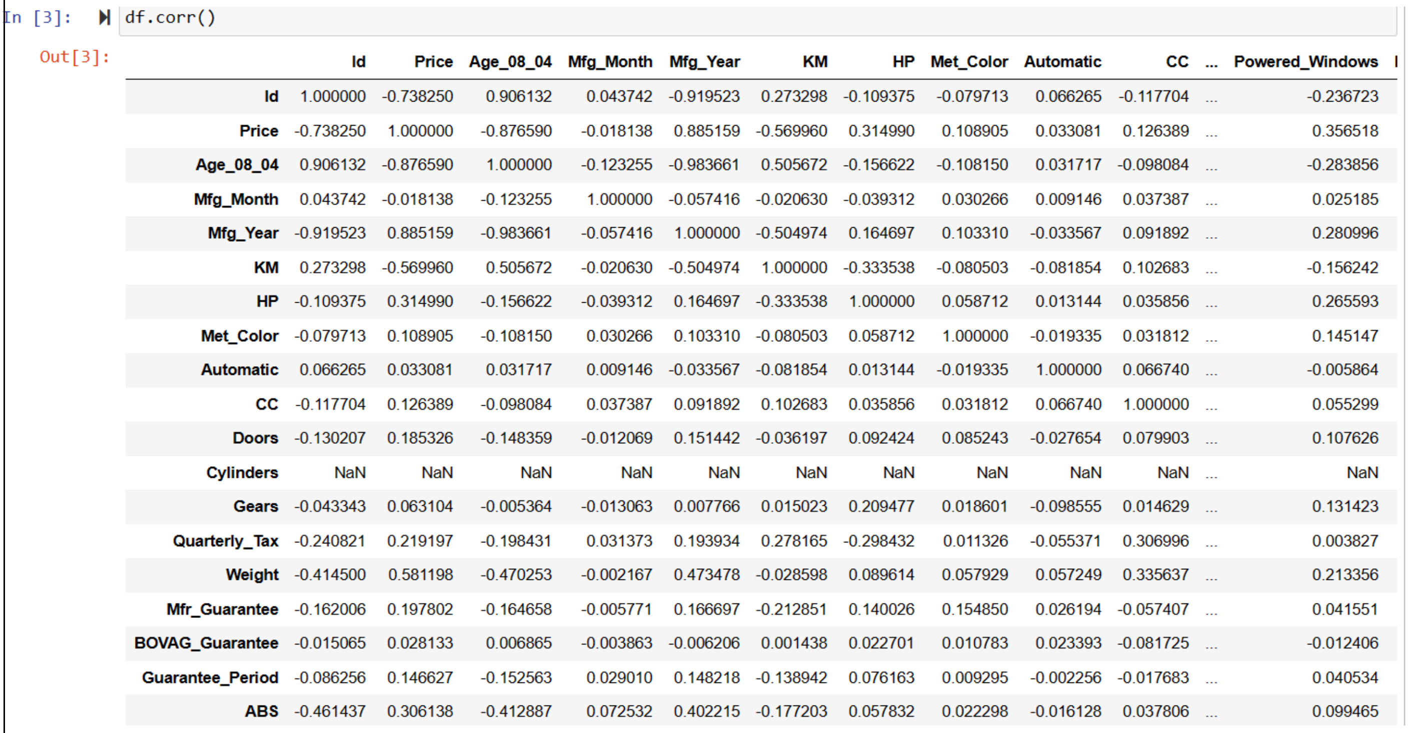
Null hypothesis H_0:There is a no significant linear relationship between the independent and dependent variable, that is, The regression coefficient value is equal to zero.

Alternative hypothesis H_1: There is a significant linear relationship between the independent and dependent variable, that is, the correlation coefficient value is not equal to zero.

- # Analysis and Interpretation :

Correlation Matrix:

Strength of Correlation:
The closer r is to 1 or -1,the stronger the correlation.
If| r| <0.3, the correlation is considered weak.
If 0.3≤| r| <0.7, the correlation is considered moderate.
If| r| ≥0.7, the correlation is considered strong.



```
In [3]:  df.corr()
```

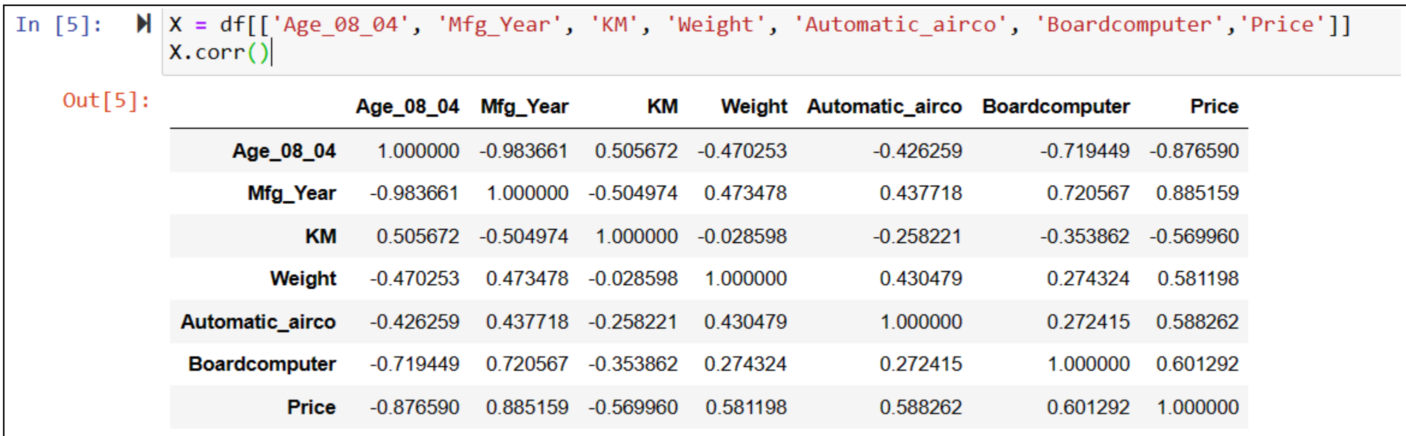
Out[3]:

	Id	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	HP	Met_Color	Automatic	CC	...	Powered_Windows
Id	1.000000	-0.738250	0.906132	0.043742	-0.919523	0.273298	-0.109375	-0.079713	0.066265	-0.117704	...	-0.236723
Price	-0.738250	1.000000	-0.876590	-0.018138	0.885159	-0.569960	0.314990	0.108905	0.033081	0.126389	...	0.356518
Age_08_04	0.906132	-0.876590	1.000000	-0.123255	-0.983661	0.505672	-0.156622	-0.108150	0.031717	-0.098084	...	-0.283856
Mfg_Month	0.043742	-0.018138	-0.123255	1.000000	-0.057416	-0.020630	-0.039312	0.030266	0.009146	0.037387	...	0.025185
Mfg_Year	-0.919523	0.885159	-0.983661	-0.057416	1.000000	-0.504974	0.164697	0.103310	-0.033567	0.091892	...	0.280996
KM	0.273298	-0.569960	0.505672	-0.020630	-0.504974	1.000000	-0.333538	-0.080503	-0.081854	0.102683	...	-0.156242
HP	-0.109375	0.314990	-0.156622	-0.039312	0.164697	-0.333538	1.000000	0.058712	0.013144	0.035856	...	0.265593
Met_Color	-0.079713	0.108905	-0.108150	0.030266	0.103310	-0.080503	0.058712	1.000000	-0.019335	0.031812	...	0.145147
Automatic	0.066265	0.033081	0.031717	0.009146	-0.033567	-0.081854	0.013144	-0.019335	1.000000	0.066740	...	-0.005864
CC	-0.117704	0.126389	-0.098084	0.037387	0.091892	0.102683	0.035856	0.031812	0.066740	1.000000	...	0.055299
Doors	-0.130207	0.185326	-0.148359	-0.012069	0.151442	-0.036197	0.092424	0.085243	-0.027654	0.079903	...	0.107626
Cylinders	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Gears	-0.043343	0.063104	-0.005364	-0.013063	0.007766	0.015023	0.209477	0.018601	-0.098555	0.014629	...	0.131423
Quarterly_Tax	-0.240821	0.219197	-0.198431	0.031373	0.193934	0.278165	-0.298432	0.011326	-0.055371	0.306996	...	0.003827
Weight	-0.414500	0.581198	-0.470253	-0.002167	0.473478	-0.028598	0.089614	0.057929	0.057249	0.335637	...	0.213356
Mfr_Guarantee	-0.162006	0.197802	-0.164658	-0.005771	0.166697	-0.212851	0.140026	0.154850	0.026194	-0.057407	...	0.041551
BOVAG_Guarantee	-0.015065	0.028133	0.006865	-0.003863	-0.006206	0.001438	0.022701	0.010783	0.023393	-0.081725	...	-0.012406
Guarantee_Period	-0.086256	0.146627	-0.152563	0.029010	0.148218	-0.138942	0.076163	0.009295	-0.002256	-0.017683	...	0.040534
ABS	-0.461437	0.306138	-0.412887	0.072532	0.402215	-0.177203	0.057832	0.022298	-0.016128	0.037806	...	0.099465

Here, Age_08_04, Mfg_Year, KM, Weight, Automatic_airco, Boardcomputer we have considered these variables as independent variables as they have a good correlation with the dependent variable(Price).

	Age_08_04	Mfg_Year	KM	Weight	Automatic_airco	Broadcomputer
Price	-0.876590	0.885159	-0.5699	0.58119	0.588262	0.601292

The correlation matrix of independent variables with respect to dependent variable:

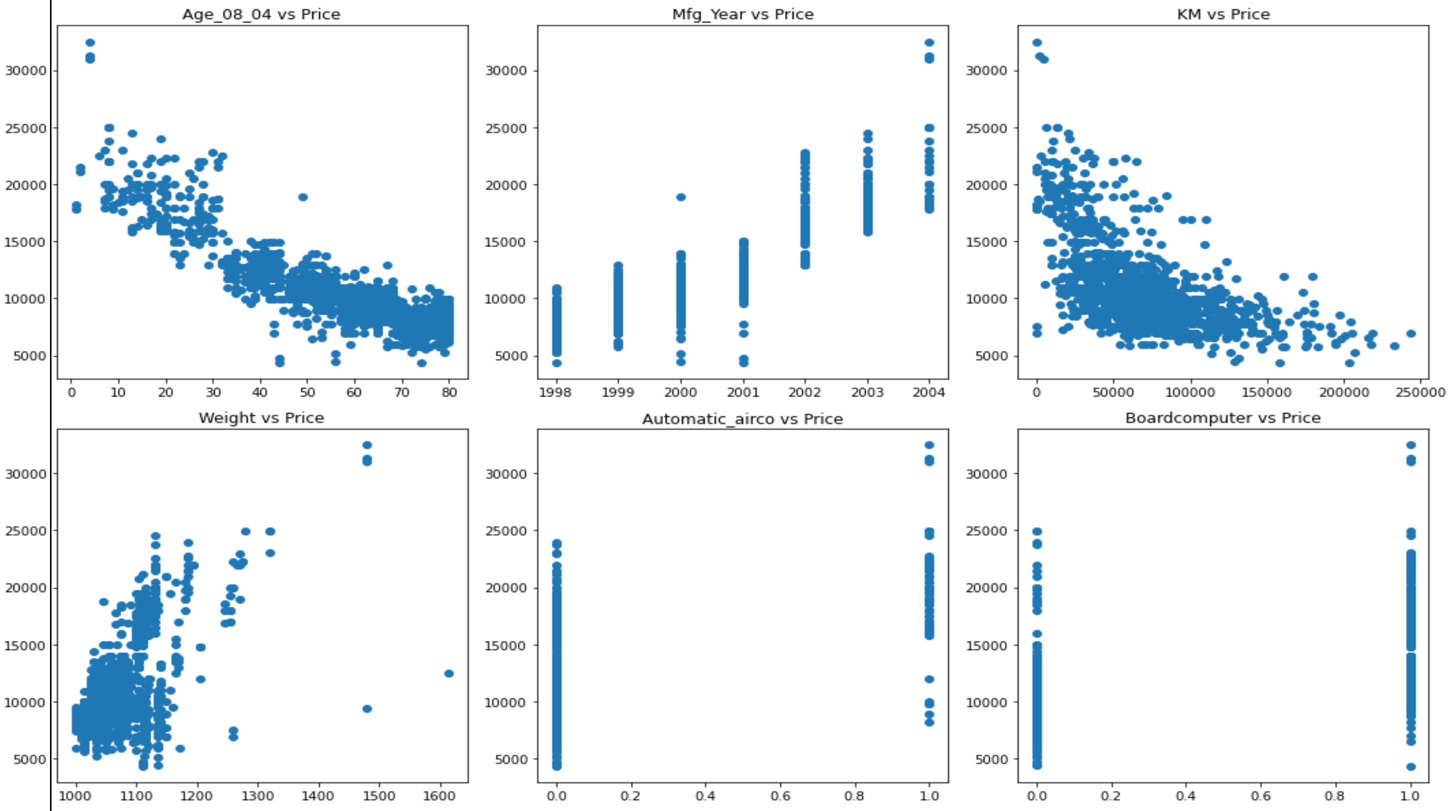


```
In [5]:  X = df[['Age_08_04', 'Mfg_Year', 'KM', 'Weight', 'Automatic_airco', 'Boardcomputer','Price']]
         X.corr()
```

Out[5]:

	Age_08_04	Mfg_Year	KM	Weight	Automatic_airco	Boardcomputer	Price
Age_08_04	1.000000	-0.983661	0.505672	-0.470253	-0.426259	-0.719449	-0.876590
Mfg_Year	-0.983661	1.000000	-0.504974	0.473478	0.437718	0.720567	0.885159
KM	0.505672	-0.504974	1.000000	-0.028598	-0.258221	-0.353862	-0.569960
Weight	-0.470253	0.473478	-0.028598	1.000000	0.430479	0.274324	0.581198
Automatic_airco	-0.426259	0.437718	-0.258221	0.430479	1.000000	0.272415	0.588262
Boardcomputer	-0.719449	0.720567	-0.353862	0.274324	0.272415	1.000000	0.601292
Price	-0.876590	0.885159	-0.569960	0.581198	0.588262	0.601292	1.000000

Scatter Plot:



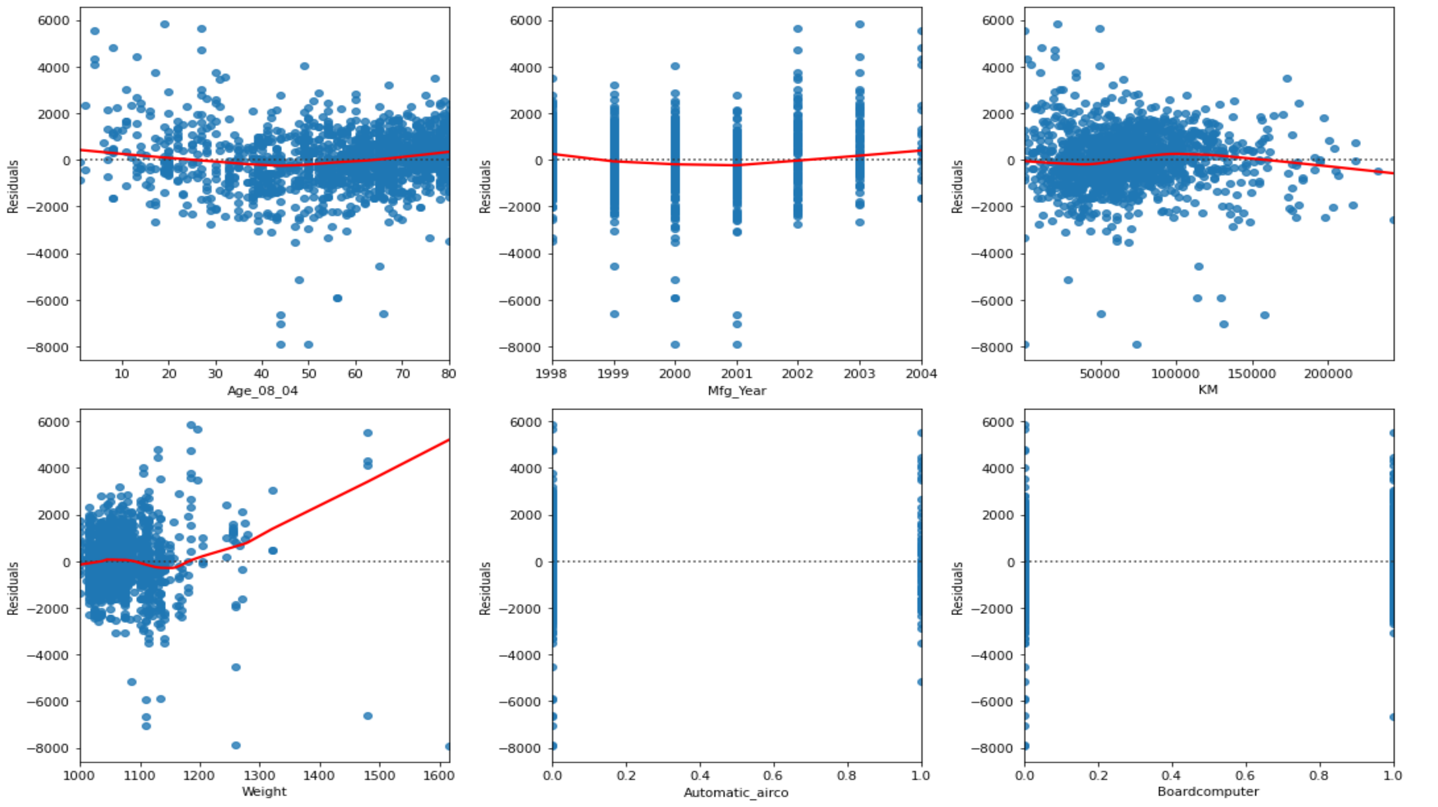
Price and Age_08_04 have scatter plot displaying a negative linear relationship as the points is scattered along a diagonal line sloping downwards from left to right.

Price and Mfg_year have scatter plot displaying a positive linear relationship as the points is scattered sloping upward from left to right.

Price and Weight have scatter plot displaying a cluster i.e. Points on the plot form distinct groups or clusters, suggesting the presence of different sub populations or categories within the data.

Price and automatic_airco scatter plot represents binary graph as automatic_airco have a binary data representing 0 as no and 1 as yes. This plot shows that car having 0 aircooler is less in price and car having 1 aircooler is high in price. Same applies for Boardcomputer vs price plot.

RESIDUAL PLOTS:



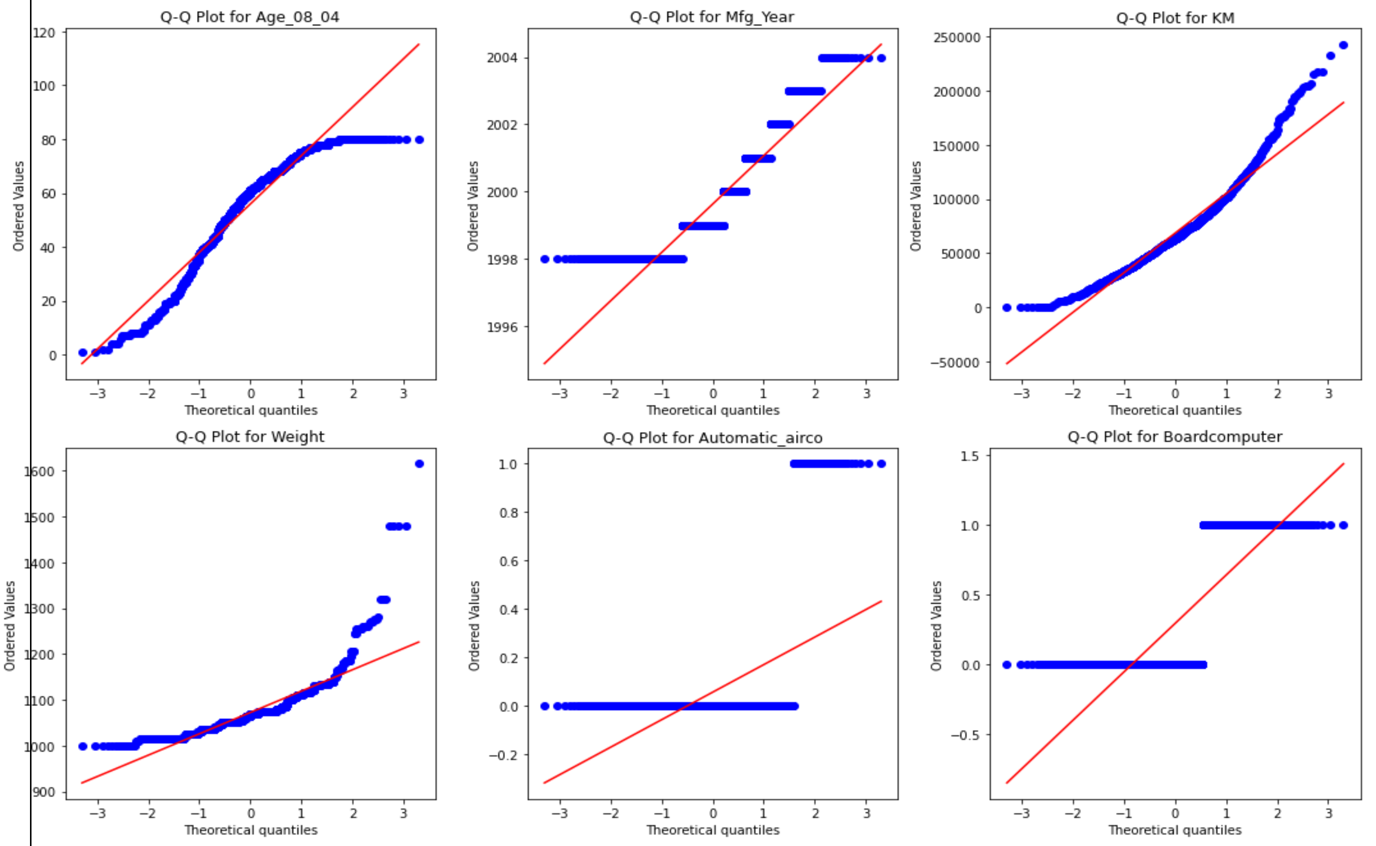
Residual plot of Age_08_04 follows linearity which implies that the residuals form a roughly horizontal band with no systematic curvature. This suggests that the relationship between the predictors and the outcome variable is adequately captured by the linear model.

Residual plot of Mfg_year follows linearity which implies that the residuals form a roughly horizontal band with no systematic curvature. This suggests that the relationship between the predictors and the outcome variable is adequately captured by the linear model.

Residual plot of KM follows linearity which implies that the residuals form a roughly horizontal band with no systematic curvature. This suggests that the relationship between the predictors and the outcome variable is adequately captured by the linear model.

Residual plot of Automatic_airco and Boardcomputer having points on 0 and 1 as it is binary data showcasing no and yes respectively.

Q-Q Plot :

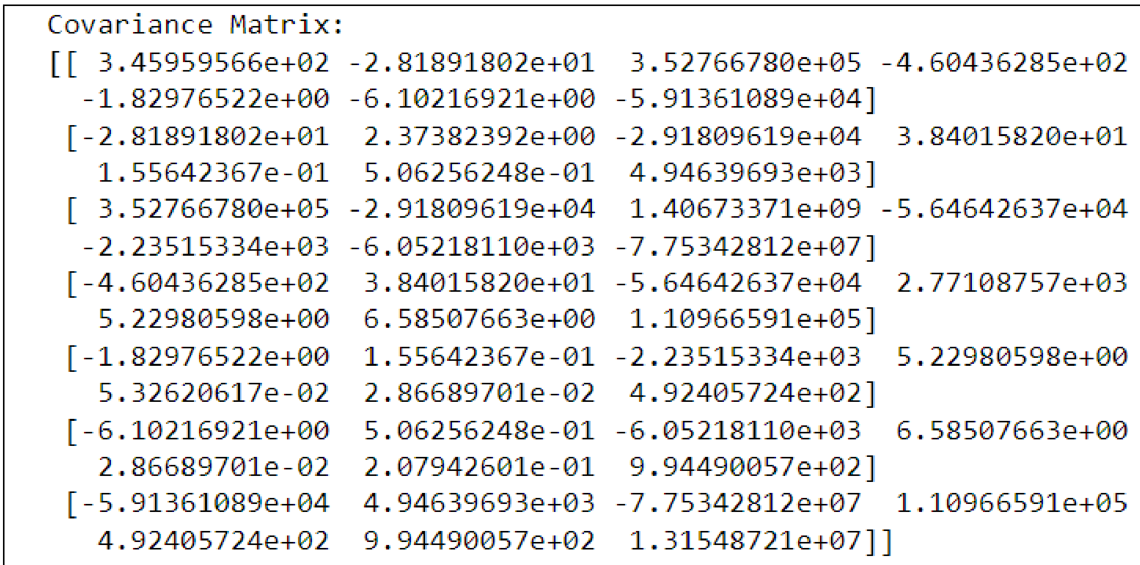


The plot for Age _08_04, KM and Weight are points closely following a straight line, it suggests that the dataset is approximately normally distributed.

The plot for Mfg_year the points fall exactly along a straight diagonal line, it indicates a perfect fit to a uniform distribution. This means that the observed quantiles match the expected quantiles of a uniform distribution, suggesting that the dataset is uniformly distributed.

QQ plot of Automatic_airco and Boardcomputer having points on 0 and 1 as it is binary data showcasing no and yes respectively.

- **Covariance matrix:**



```
Covariance Matrix:
[[ 3.45959566e+02 -2.81891802e+01  3.52766780e+05 -4.60436285e+02
  -1.82976522e+00 -6.10216921e+00 -5.91361089e+04]
 [-2.81891802e+01  2.37382392e+00 -2.91809619e+04  3.84015820e+01
   1.55642367e-01  5.06256248e-01  4.94639693e+03]
 [ 3.52766780e+05 -2.91809619e+04  1.40673371e+09 -5.64642637e+04
  -2.23515334e+03 -6.05218110e+03 -7.75342812e+07]
 [-4.60436285e+02  3.84015820e+01 -5.64642637e+04  2.77108757e+03
   5.22980598e+00  6.58507663e+00  1.10966591e+05]
 [-1.82976522e+00  1.55642367e-01 -2.23515334e+03  5.22980598e+00
   5.32620617e-02  2.86689701e-02  4.92405724e+02]
 [-6.10216921e+00  5.06256248e-01 -6.05218110e+03  6.58507663e+00
   2.86689701e-02  2.07942601e-01  9.94490057e+02]
 [-5.91361089e+04  4.94639693e+03 -7.75342812e+07  1.10966591e+05
   4.92405724e+02  9.94490057e+02  1.31548721e+07]]
```

Certainly! Let's interpret the covariance values with respect to the target variable, 'Price':

1. *Age_08_04*:
- Covariance with Price: 3.45959566e+02
- Positive covariance indicates that as the age of the car increases, the price tends to increase. This suggests that newer cars tend to have higher prices, which is a common expectation.

2. *Mfg_Year*:
- Covariance with Price: 2.37382392e+00
- Positive covariance also indicates that as the manufacturing year of the car increases, the price tends to increase. This aligns with the interpretation of 'Age_08_04', as newer cars typically have higher prices.

3. *KM* (Kilometers driven):
- Covariance with Price: 3.52766780e+05
- A large positive covariance suggests that as the number of kilometers driven increases, the price tends to decrease. This implies that cars with higher mileage generally have lower prices.

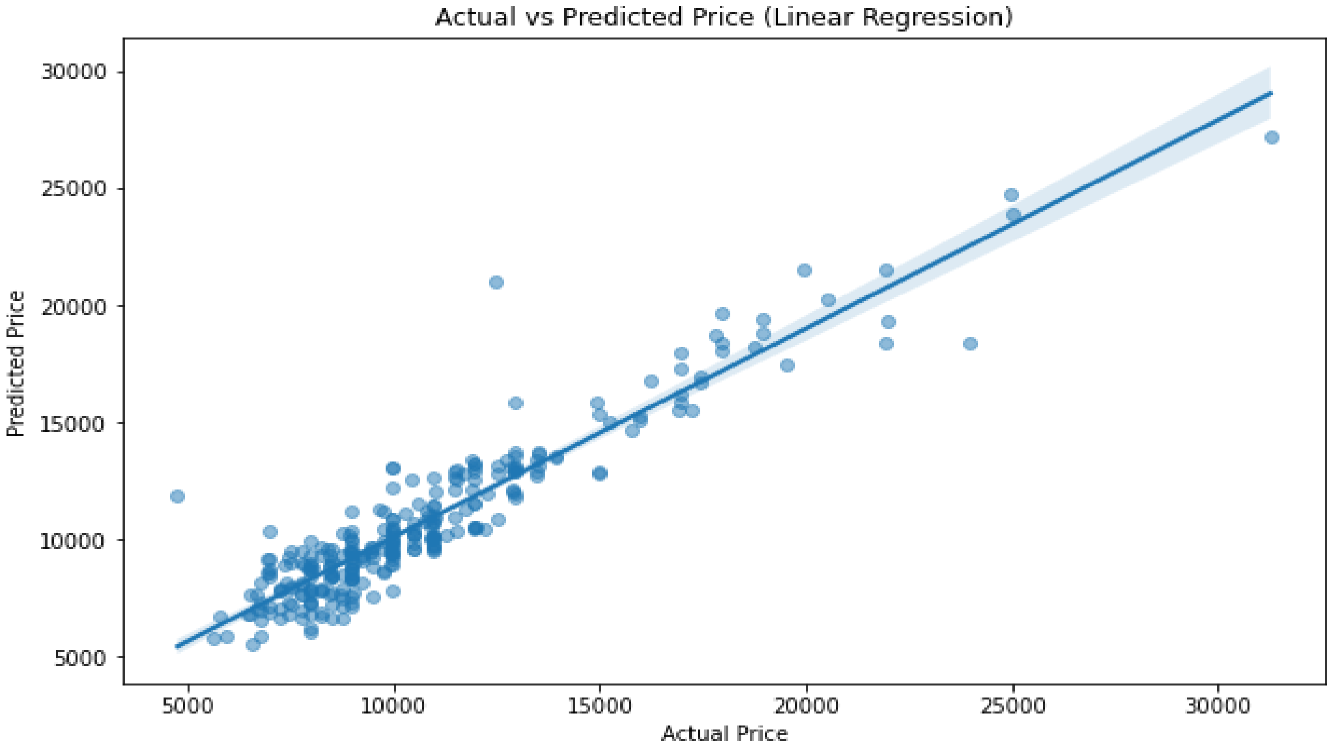
4. *Weight*:
- Covariance with Price: -4.60436285e+02
- Negative covariance indicates that as the weight of the car increases, the price tends to decrease. This might seem counterintuitive, but it suggests that heavier cars might not necessarily command higher prices in this dataset.

5. *Automatic_airco* (Presence of Automatic Air Conditioning):
- Covariance with Price: -1.82976522e+00
- Negative covariance suggests that cars with automatic air conditioning tend to have slightly lower prices. This could be due to various factors such as market demand and vehicle type.

6. *Boardcomputer*:
- Covariance with Price: -6.10216921e+00
- Negative covariance indicates that cars equipped with a board computer tend to have slightly lower prices. Again, this could be influenced by market factors and consumer preferences.

These interpretations provide insights into how each feature relates to the target variable, 'Price', based on their covariance values in the dataset.

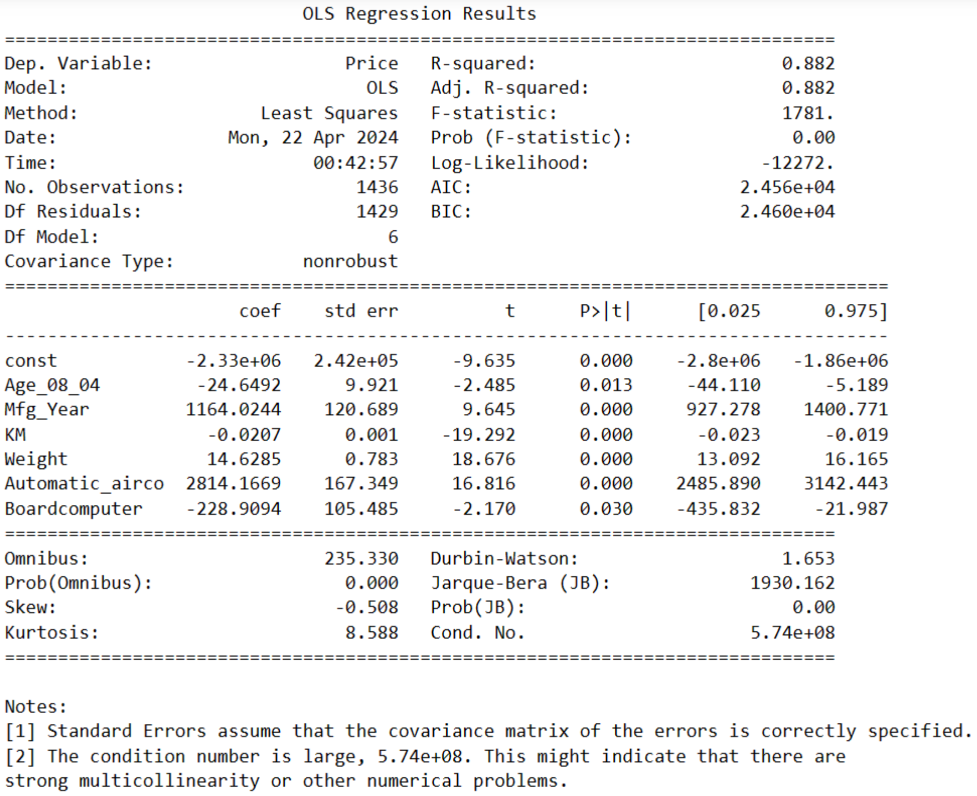
● FITTING REGRESSION Model:



1. **Direction of the Line**: The direction of the line indicates whether there's a positive or negative relationship between the variables. If the line slopes upwards from left to right, it indicates a positive relationship, meaning that as one variable increases, the other variable tends to increase as well. Conversely, if the line slopes downwards, it indicates a negative relationship.

2. **Strength of the Relationship**: The closeness of the data points to the trend line indicates the strength of the relationship between the variables. If the data points are tightly clustered around the line, it suggests a strong relationship, whereas if the points are more scattered, the relationship may be weaker.

● MODEL SUMMARY



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.882
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     1781.
Date:                Mon, 22 Apr 2024   Prob (F-statistic):               0.00
Time:                        00:42:57   Log-Likelihood:                -12272.
No. Observations:                1436   AIC:                         2.456e+04
Df Residuals:                    1429   BIC:                         2.460e+04
Df Model:                           6
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const             -2.33e+06   2.42e+05     -9.635      0.000    -2.8e+06   -1.86e+06
Age_08_04          -24.6492      9.921     -2.485      0.013     -44.110      -5.189
Mfg_Year          1164.0244    120.689      9.645      0.000     927.278    1400.771
KM                  -0.0207      0.001    -19.292      0.000      -0.023      -0.019
Weight              14.6285      0.783     18.676      0.000      13.092      16.165
Automatic_airco   2814.1669    167.349     16.816      0.000    2485.890    3142.443
Boardcomputer     -228.9094    105.485     -2.170      0.030    -435.832     -21.987
==============================================================================
Omnibus:                      235.330   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1930.162
Skew:                          -0.508   Prob(JB):                         0.00
Kurtosis:                       8.588   Cond. No.                     5.74e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.74e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Multiple R	0.939159607
R Square	0.882020767
Adjusted R Square	0.881525403
Standard Error	1248.406251
Observations	1436

Strong correlation (Multiple R) of approximately 0.939 between predictors and the dependent variable.

Goodness-of-fit (R Square) of around 88.2%, suggesting that the model explains a significant portion of the variability in the dependent variable.

Adjusted R Square is nearly identical to R Square, indicating that adding more predictors does not notably improve the model's explanatory power.

Standard Error of the Estimate is approximately 1248.41, reflecting the average deviation of observed values from predicted values.

The analysis includes 1436 observations.

Overall, the model exhibits a robust fit to the data, explaining a substantial portion of the variance in the dependent variable with strong correlation, low error, and consistency across adjustments for predictor inclusion.

Hypothesis Testing:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-2329862.943	241805.5701	-9.635274083	2.48E-21	-2804194.905	-1855530.98	-2804194.905	-1855530.98
Age_08_04	-24.64922168	9.920592728	-2.484652113	0.013081572	-44.10970895	-5.188734416	-44.10970895	-5.188734416
Mfg_Year	1164.024386	120.6890858	9.644818978	2.28E-21	927.2776021	1400.771169	927.2776021	1400.771169
KM	-0.020715235	0.001073766	-19.2921317	6.91E-74	-0.022821562	-0.018608909	-0.022821562	-0.018608909
Weight	14.6285472	0.783285394	18.67588405	8.09E-70	13.09203463	16.16505977	13.09203463	16.16505977
Automatic_airco	2814.166881	167.3492139	16.81613445	4.85E-58	2485.890402	3142.44336	2485.890402	3142.44336
Boardcomputer	-228.9093664	105.4852373	-2.170060686	0.030166535	-435.8318935	-21.98683934	-435.8318935	-21.98683934

Intercept: The intercept coefficient indicates the estimated value of the dependent variable when all predictor variables are zero. Here, it's approximately -2,329,863. The p-value (2.48E-21) is very low, suggesting that the intercept is significantly different from zero.

Age_08_04: For every unit increase in "Age_08_04" (presumably age of the vehicle), the dependent variable decreases by approximately 24.65 units. The p-value (0.0131) suggests that this effect is statistically significant.

Mfg_Year: For every unit increase in "Mfg_Year" (manufacturing year of the vehicle), the dependent variable increases by approximately 1164.02 units. The p-value (2.28E-21) indicates statistical significance.

KM: For every unit increase in "KM" (kilometers driven), the dependent variable decreases by approximately 0.0207 units. The p-value (6.91E-74) suggests this effect is highly significant.
Weight: For every unit increase in "Weight" of the vehicle, the dependent variable increases by approximately 14.63 units. The p-value (8.09E-70) indicates statistical significance.

Automatic_airco: Vehicles with automatic air conditioning have, on average, a dependent variable approximately 2814.17 units higher than those without. The p-value (4.85E-58) suggests this effect is highly significant.

Boardcomputer: Vehicles equipped with a board computer have, on average, a dependent variable approximately 228.91 units lower than those without. The p-value (0.0302) suggests this effect is statistically significant.
Overall, the interpretation suggests that these predictor variables have statistically significant effects on the dependent variable, as indicated by their coefficients and p-values.

ANOVA					
	df	SS	MS	F	Significance F
Regression	6	16650119001	2775019834	1780.550198	0
Residual	1429	2227122462	1558518.168		
Total	1435	18877241464			

Regression: The regression model, with 6 degrees of freedom (df), explains a significant amount of the variation in the dependent variable, as indicated by the high F-statistic of 1780.55 and a very low p-value of 0. This suggests that the model's predictors collectively have a strong influence on the dependent variable.

Residual: The residual error, representing unexplained variability after accounting for the regression model, has 1429 degrees of freedom. The mean square (MS) residual is 1558518.168.

Total: The total variation in the dependent variable across all observations is summarized, with a total of 1435 degrees of freedom.

Overall, the results indicate that the regression model is highly significant in explaining the variability in the dependent variable, with very low probability of obtaining such results by chance (p-value = 0).