# BNM854 DESCRIPTIVE ANALYTICS COURSEWORK

# BUSINESS REPORT: VISUALISATION AND STATISTICAL ANALYSIS OF A REAL WORLD DATA SET

**STUDENT NAME – GAURAV RANDVIE**

**STUDENT NUMBER - 230376085**

# Table of Contents

## 1. INTRODUCTION

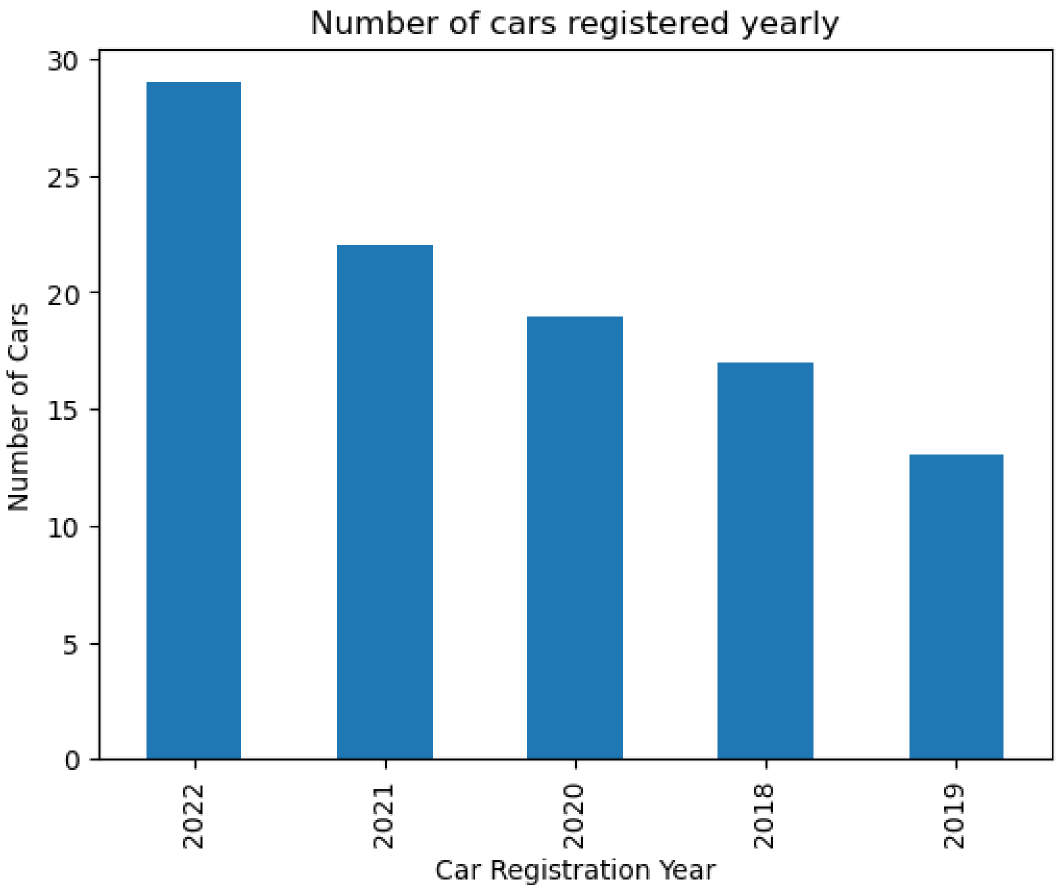
We are examining online car data, with a particular emphasis on the "TOYOTA HILUX" between 2018 and 2022. By excluding 2023, any outliers brought on by the addition of a new variety are avoided.

The website www.autotrader.co.uk provided the data, which presented difficulties in a number of forms during collection. Regression analysis and visual aids required preprocessing, which included converting 'Engine Size' from a text including BHP or PS to an integer format.

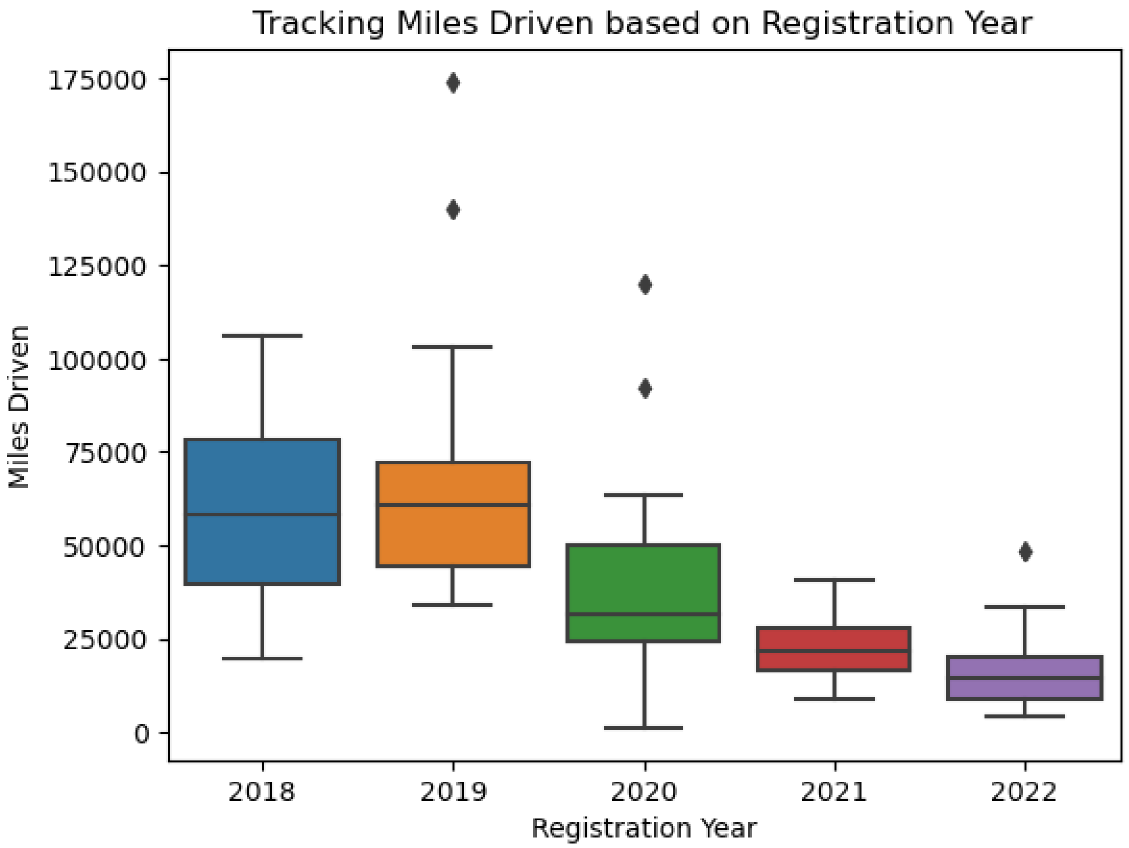
Our goal is to comprehend the variables affecting the cost of second hand cars. We created models to calculate car expenses using separate variables. Testing the validity of the model's predictions is part of validation.

The 5-year span guarantees stability in model years, popularity, and technology, which improves our capacity to evaluate the market. This strategy avoids unfair comparisons seen in a decade's worth of data, where later models might exhibit advancements in technology and fuel efficiency.

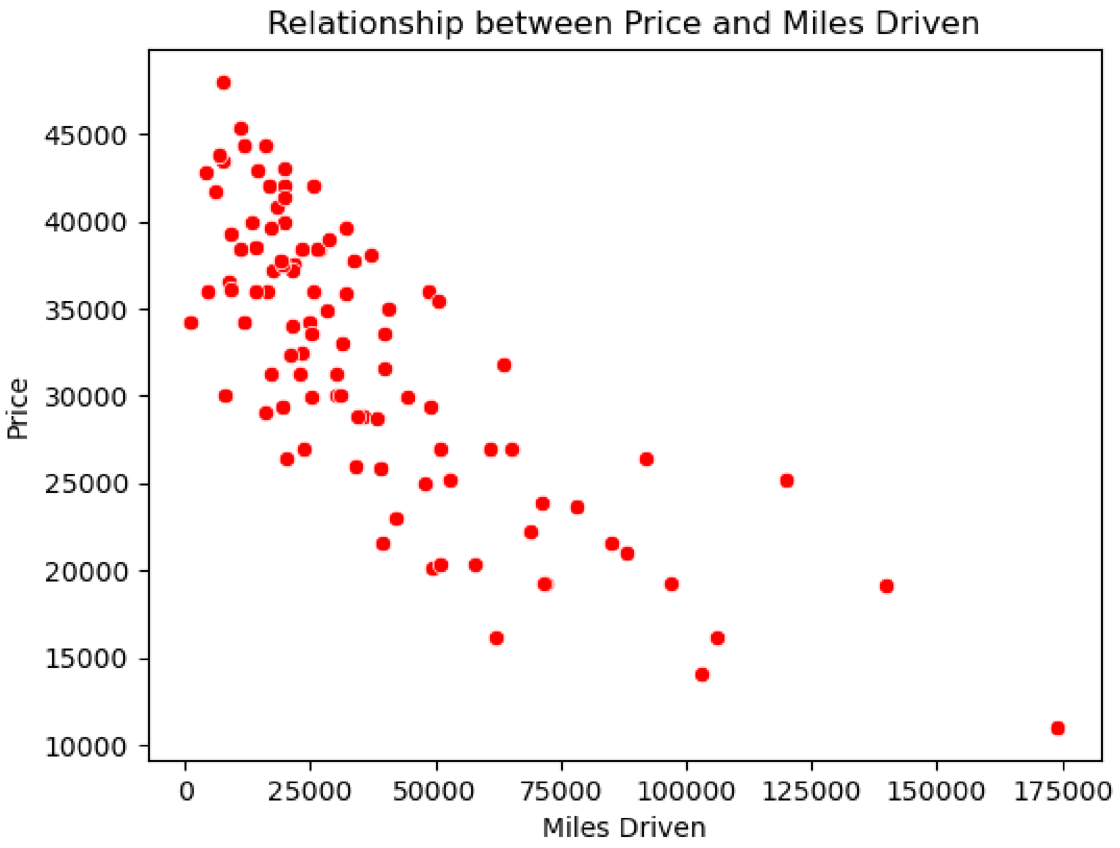
## 2. DATA VISUALIZATION



To examine the number of cars registered in the last few years within our sample of cars, I'm creating a bar plot. We can examine the years with the highest number of registered autos by doing this. We can determine that, with 29 cars registered, 2022 has the largest number, followed by 2021 and 2020. With 13 registered cars, 2019 has the fewest number of cars registered.



We examine the miles driven by registered automobiles annually using a boxplot. "Miles Driven" is on the y-axis, while "Registration Year" is on the x-axis. The mileage of older autos is often greater. Interestingly, the median mileage of automobiles in 2018 was over 65,000, which is just less than the median mileage of 67,000 in 2019. For automobiles registered in 2020, 2021, and 2022, the median mileage has consistently decreased. This pattern is consistent with the hypothesis that fewer newer automobiles than those registered in prior years are being driven.



The scatterplot shows the distribution of the number of miles driven and the price of a car. The scatter plot shows a negative correlation between the two variables, meaning that as the number of miles driven increases, the price of the car decreases.

This is because cars depreciate over time, meaning that they lose value as they are used more. This is due to a number of factors, such as wear and tear, the age of the car, and the overall condition of the car.

## 3. DATA SUMMARY

**SUMMARIZING THE DATA**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Price** | 100 | 32393.97 | 8103.81 | 10995 | 26845.50 | 33802 | 38400 | 47994 |
| **Miles Driven** | 100 | 36873.43 | 30500.62 | 1200 | 17171.75 | 26529 | 48775 | 174000 |
| **Engine Power** | 100 | 172.08 | 26.94 | 146 | 150 | 150 | 204 | 208 |
| **Reg Year** | 100 | 2020 | 1.44 | 2018 | 2019 | 2021 | 2022 | 2022 |

When we find the summary of the data, we get the table given above. Which depicts crucial information about the data sample:

Price:

- The dataset consists of 100 observations related to the prices of vehicles.
- The mean (average) price is £32,393.97, with a standard deviation of £8,103.81, indicating the dispersion of prices around the mean.
- Prices range from a minimum of £10,995 to a maximum of £47,994.
- The interquartile range (IQR) is from £26,845.50 to £38,400, capturing the middle 50% of the data.

Miles Driven:

- This variable represents the miles driven by the vehicles.
- The mean miles driven is 36,873.43, with a standard deviation of 30,500.62.
- The range is from a minimum of 1,200 miles to a maximum of 174,000 miles.
- The IQR is from 17,171.75 miles to 48,775 miles.

Engine Power:

- The dataset includes information about the engine power of the vehicles.
- The mean engine power is 172.08, with a standard deviation of 26.94.
- Engine power ranges from a minimum of 146 to a maximum of 208.
- The majority of the data (from the 25th to the 75th percentile) is between 150 and 204.

Registration Year:

- This variable represents the registration year of the vehicles.
- The mean registration year is 2020, with a standard deviation of 1.44 years.
- The dataset includes vehicles registered from 2018 to 2022.
- The interquartile range for registration years is from 2019 to 2022**.**

**COMPARING PRICE AND MILES DRIVEN**

| Price | Miles Driven | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|
| | **0-50k** | 77 | 35418.45 | 6019.54 | 20100 | 31194.0 | 35988 | 39588.00 | 47994 |
| | **50k-100k** | 18 | 23703.83 | 4838.10 | 16188 | 20395.5 | 22948 | 26846.25 | 35394 |
| | **100k+** | 5 | 17101.40 | 5102.09 | 10995 | 14050.0 | 16194 | 19080.00 | 25188 |

**Insights on the table above:**

- Generally, there is a relationship between miles driven and price, which aligns with typical depreciation patterns.
- Vehicles with lower miles driven (0-50k) tend to have higher average prices compared to those with higher driven distances.
- The standard deviation in prices is relatively smaller for the 0-50k miles range, indicating more consistency in pricing for lesser-distance-driven vehicles.
- As the driven distance increases, the average price decreases, and the variability in prices tends to increase.

## 4. CONFIDENCE INTERVAL

Assuming normal distribution in used car prices, I've calculated a 95% confidence interval for the mean price. This interval indicates a range within which we are 95% confident the true population mean is situated.

**Steps to find the confidence level:**

1. Calculate the mean and standard deviation of the 'Price' column.

2. Determine the sample size.
3. Choose the desired confidence levels (i.e. 95%)
4. Calculate the standard error of the mean using the Z-score for the chosen confidence levels.
5. Compute the confidence intervals using the formula: mean ± (Z * (std_dev / sqrt(sample_size))).

**Result:**

| Confidence Level | Lower Bound | Upper Bound |
|---|---|---|
| 95 | 30805.62324 | 33982.31676 |

**Interpretation of the result:**

- The result shows the lower and the upper bound, i.e. 30805.62324 and 33982.31676 respectively.
- This means that we are 95% confident that the population mean is going to be somewhere in between the lower bound and the upper bound.

### 5. HYPOTHESIS TESTING

As I don't have data from different locations, therefore, I have considered the average price of the Toyota Hilux mentioned in the official website https://www.cargurus.co.uk/Cars/l-Used-Toyota-Hi-Lux-d2355. Then, I will consider the average price of the sample with the average price that I have found on the website. I have collected sample prices for Hilux depending on the distance driven.

| Distance Driven | Average Price |
|---|---|
| 0-20K | 38016 |
| 20K-40K | 34848 |
| 40K-60K | 29990 |

We will need to perform hypothesis testing to check if the mean has any significant difference or not. To do so, we will first have to perform the following steps:

- Group the data by the distance driven (0-20k, 20k-40k,40k-60k).
- Find the **mean** price of each group or **predict** the value using a regression model.
- Compare the value that you get with the value we collected from the website / Hypothesis Testing.

**Results:**

| Hypothesis Test for 0-20k Miles Range | |
|---|---|
| Test Statistics: | 1.6480 |
| P-value: | 0.1091 |
| Result: | Fail to reject the null hypothesis: There is no significant difference in the mean. |

| Hypothesis Test for 20k-40k Miles Range | |
|---|---|
| Test Statistics: | -1.7856 |
| P-value: | 0.0828 |
| Result: | Fail to reject the null hypothesis: There is no significant difference in the mean. |

| Hypothesis Test for 40k-60k Miles Range | |
|---|---|
| Test Statistics: | -1.4552 |
| P-value: | 0.1712 |
| Result: | Fail to reject the null hypothesis: There is no significant difference in the mean. |

**Interpretation of the results:**

We performed a hypothesis test and we realized that the means of all the ranges don’t show any significant difference from the mean of the sample collected from the website.

Therefore, we reject the null hypothesis. That means our model is a good fit as it predicts a value that is proven to be a legitimate value when tested.

## 6. CORRELATION OF DATA

To find the correlation between the variables we can simply use the corr() function. This function gives us the Correlation matrix showing the relationship between variables.

**Correlation Matrix:**

| | **Price** | **Miles Driven** | **Registration Year** | **Engine Power** |
|---|---|---|---|---|
| **Price** | 1 | -0.77 | 0.84 | 0.69 |
| **Miles Driven** | -0.77 | 1 | -0.62 | -0.39 |
| **Registration Year** | 0.84 | -0.62 | 1 | 0.68 |
| **Engine Power** | 0.69 | -0.39 | 0.68 | 1 |

**Interpretation of the above correlation matrix:**

- The above matrix defines the correlation of each variable with each other. Value = 1 represents that the compared variables are the same.
- Values above 0.7 represent a strong positive correlation and values below -0.7 represent the strong negative correlation between the variables.
- In this case, Price and Miles show a negative correlation, while Registration Year and Price show a strong positive correlation.
- Miles and registration year show a moderate negative correlation.
- Miles and Engine Power depict a moderate negative correlation.
- Registration Year and Engine Power show a moderate positive correlation.

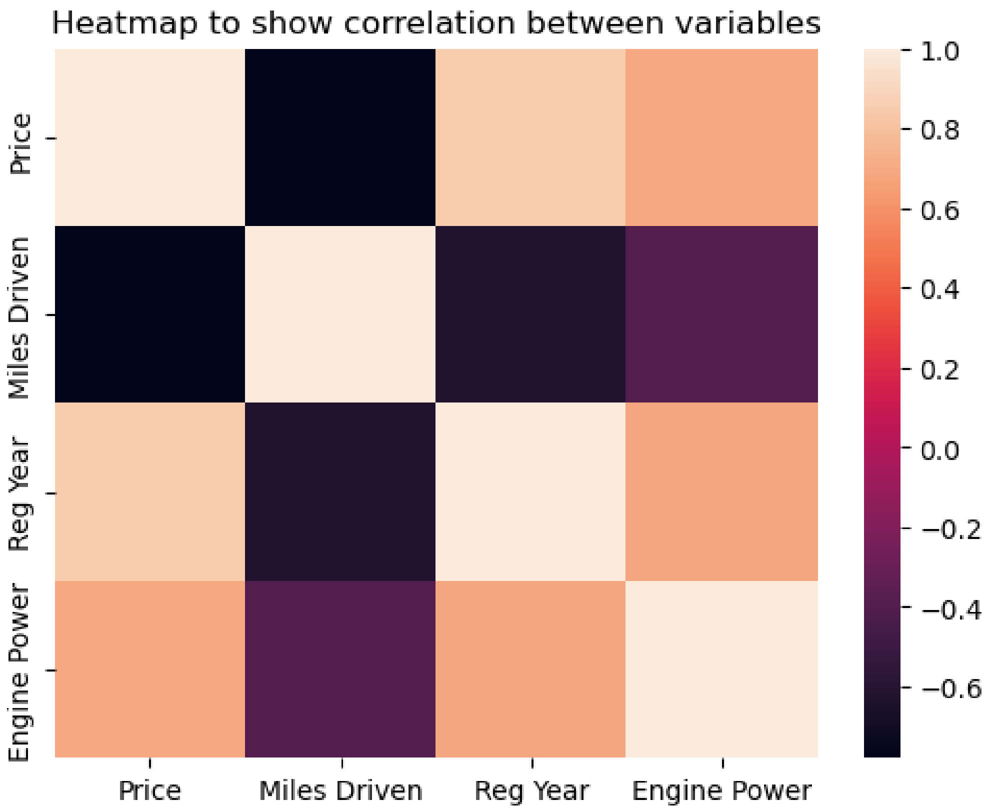
**Correlation with Price variable:**

| **Price** | **Miles Driven** | **Registration Year** | **Engine Power** |
|---|---|---|---|
| | **-0.77** | **0.84** | **0.69** |

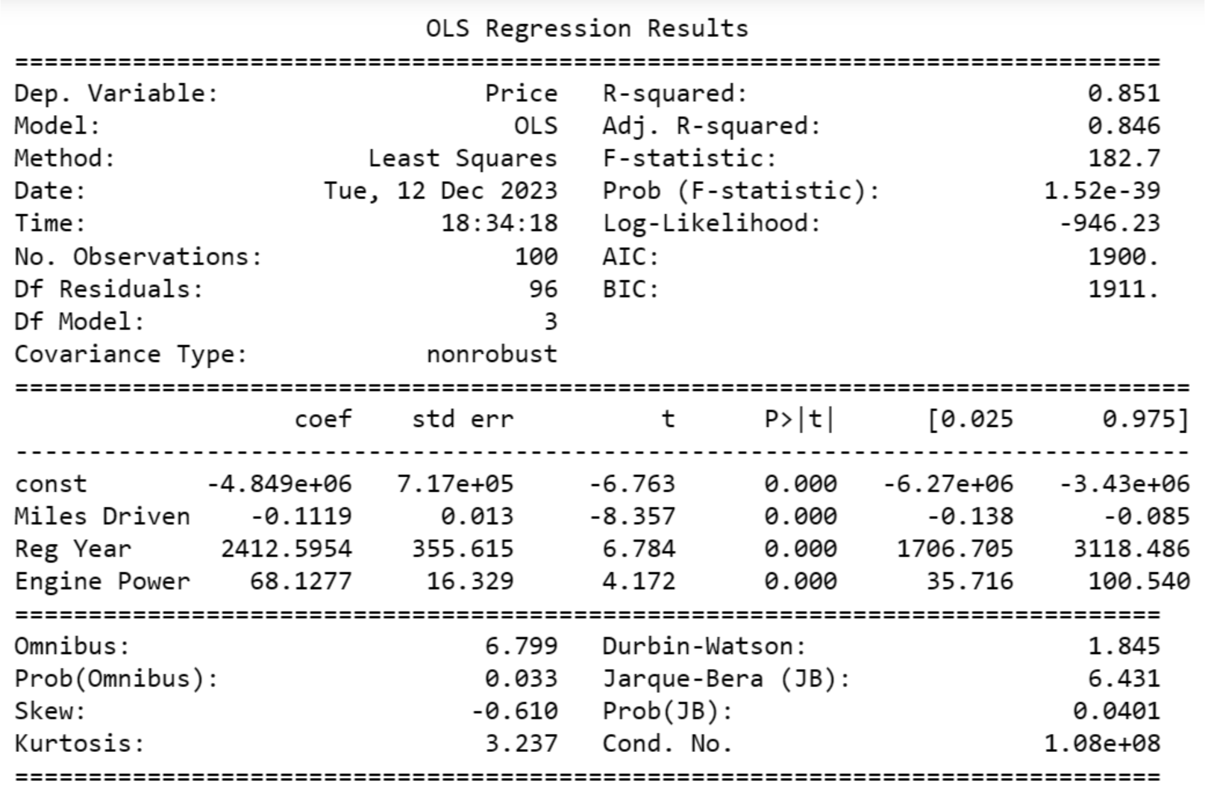
**Interpretation of the above table:**

When we find a correlation between the dependent variable (Price) with the other independent variables (Engine Power, Registration Year, Miles Driven), we can analyze that:

- Price has a strong negative correlation with Miles Driven, i.e. When one variable increases the other decreases.
- Price shows a strong positive correlation with the Registration Year, i.e. Price is highly impacted by the Registration Year.
- Price then shows a moderate positive correlation with Engine Power, indicating that there is a moderate tendency for them to increase together.



## 7. REGRESSION ANALYSIS



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.851
Model:                            OLS   Adj. R-squared:                  0.846
Method:                 Least Squares   F-statistic:                     182.7
Date:                Tue, 12 Dec 2023   Prob (F-statistic):           1.52e-39
Time:                        18:34:18   Log-Likelihood:                -946.23
No. Observations:                 100   AIC:                             1900.
Df Residuals:                      96   BIC:                             1911.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -4.849e+06   7.17e+05     -6.763      0.000   -6.27e+06   -3.43e+06
Miles Driven    -0.1119      0.013     -8.357      0.000      -0.138      -0.085
Reg Year      2412.5954    355.615      6.784      0.000    1706.705    3118.486
Engine Power    68.1277     16.329      4.172      0.000      35.716     100.540
==============================================================================
Omnibus:                        6.799   Durbin-Watson:                   1.845
Prob(Omnibus):                  0.033   Jarque-Bera (JB):                6.431
Skew:                          -0.610   Prob(JB):                       0.0401
Kurtosis:                       3.237   Cond. No.                     1.08e+08
==============================================================================
```

In the above OLS Regression Model, the model depicts the following details:

1. R-squared and Adjusted R-squared: The R-squared is 0.851, indicating that around 85.1% of the Price variability is explained by Miles Driven, Registration Year, and Engine Power. The adjusted R-squared, slightly lower at 0.846, considers the model's complexity, suggesting potential avoidance of overfitting with the chosen variables.
2. Coefficients: The R-squared is 0.851, indicating that around 85.1% of the Price variability is explained by Miles Driven, Registration Year, and Engine Power. The adjusted R-squared, slightly lower at 0.846, considers the model's complexity, suggesting potential avoidance of overfitting with the chosen variables.
3. Model Fit: The high R-squared value (0.851) indicates that the model explains a significant proportion of the variance in the dependent variable.

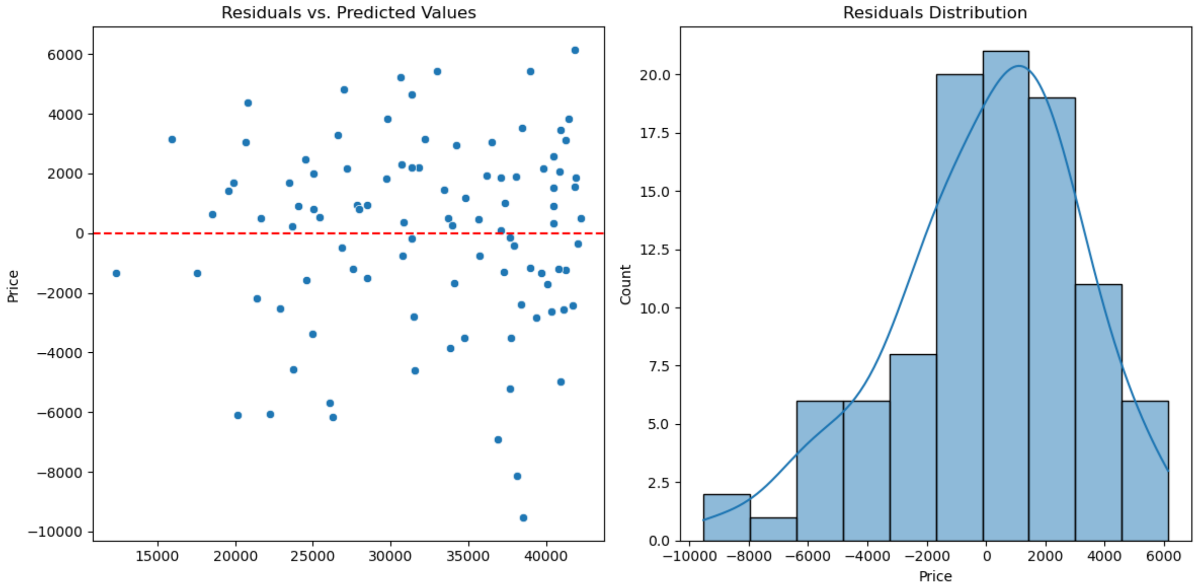
Parsimoniousness Model:

There are multiple ways to check the parsimoniousness of the model:

1. P-value : Minimum value for p-value indicates that the variable is likely to be statistically significant.
2. Adjusted R-square: The higher the value of adjusted R-square, greater is the fit of model.
3. AIC and BIC : The OLS Regression Model shows that the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are lower. AIC value being 1900 and BIC value being 1911.

We choose this model, as the adjusted R-square value is comparatively higher and the AIC/BIC score are significantly less.

## 8. RESIDUAL ANALYSIS



• Linearity: The left graph, a scatter plot of residuals vs. predicted values, shows random dispersion around the y=0 line, indicating a presumed linear relationship.

• Independence: This assumes no correlation between consecutive residuals, ensuring their independence.

• Homoscedasticity: The absence of a funnel shape in the scatter plot of residuals vs. predicted values (left graph) suggests constant variance across predictor variables.

• Normality: The right histogram's bell-shaped curve confirms the normal distribution of residuals.

• No Multicollinearity: Previous correlation testing indicates low correlation among independent variables, ensuring no significant multicollinearity concerns.

Conclusion on residual analysis: We can conclude that our model passes all the five regression analysis.

### 9. STATISTICAL MODEL

Price = β0 +β1 × Miles Driven+β2 × Registration Year + β3 × Engine Power + ϵ

Here:

- Price is the dependent variable (the predicted second-hand car price).
- β0 is the intercept term.
- β1, β2 and β3 are the coefficients for the independent variables (‘Miles Driven’, ‘ Registration Year’ and ‘Engine Power’).
- ϵ represents the error term.

Uses of the Derived Statistical Model:

• Prediction: The model predicts the dependent variable based on given independent variable values.

• Understanding Relationships: Coefficients (β) indicate the strength and direction of relationships between independent and dependent variables, with positive or negative associations.

• Inference and Hypothesis Testing: Statistical tests on coefficients enable inferences about population parameters and assess the significance of variables or variable groups.

• Model Evaluation: Evaluating goodness-of-fit measures, like R-squared, gauges how well the model explains the dependent variable's variability.

• Policy and Decision Making: The model's insights guide decision-making processes, informing strategies, resource allocation, or identifying influential factors.

**Let’s now ask the model a predict price when given the independent variables:**

Miles Driven = 30000, Registration Year= 2020, Engine Size = 150.

The statistical model looks like this:

Price = -4.849e+06 – 0.1119 x 30000(Miles Driven)+ 2412.5954 x 2020(Registration year) + 68.1277 x 150(Engine Size)

**Price = 31304.858**

So, the model predicts the car price to be 31304.858 pounds.

## 10. MODEL ANALYSIS

**To solve this problem we have used linear regression.**

**Reasons for using linear regression:**

- Interpretability: Linear regression offers a clear understanding of variable relationships. Each coefficient (β) signifies the change in the dependent variable with a one-unit change in the corresponding independent variable.
- Simplicity: Linear regression is straightforward and easy to implement, serving as an effective starting point when assuming a linear relationship between variables.
- Assumption of Linearity: Accurate predictions are achievable with linear regression when relationships between independent and dependent variables are approximately linear.
- Model Transparency: The model's transparency facilitates easy interpretation of variable impacts, aiding decision-making.
- Assumption of Independence: Linear regression assumes independence of observations, often reasonable in practical scenarios.

**Analysis of the model:**

- To make sure that the model used is working fine or not, we performed diagnostics, such as residual analysis to access the validity of the model and identify the areas for improvement.

- We used linear regression model as we knew that there is certain correlation between dependent and independent variables.
- After analysing the outputs, we can say that the data fits the model in a significant way (84%) and the model performs well when asked to predict price given the independent variables (Miles driven, Registration Year, Engine Power).