BNM854

Business report: **Visualisation and Statistical Analysis of a Real World Data Set**

Coursework Details: **2nd Hand Car Market Case Study**

Candidate Id- 749403

Executive Summary :

1. Description :

This report aims to conduct a thorough analysis of the second- hand auto market within the Leeds (LS13DA) area, considering specifically on the BMW 3 Series model. The primary goal is to identify important variables impacting the pricing of second- hand BMW 3 Series cars . The study delves into statistical modeling to estimate request values grounded on factors similar as avail, machine size, and voluntary features.

2. Source of Data:

The dataset for this study was sourced from www.autotrader.co.uk , a estimable online platform specializing in second- hand cars. A population of 300 second- hand BMW 3 Series buses within the Leeds area was precisely named for analysis. The data includes pivotal information similar as price, mileage, fuel type and voluntary features.

3. Limitations and Challenges:

Several limitations and challenges were encountered during the data collection and analysis process. These include implicit variations in data fitness across lists, the reliance on openly available information, and possible aptitudes in online registries. Also, the study is constrained by the accessibility of data within the specified Leeds area and the named car model.

4. Sampling Method: Random

The sampling system used in this study is random, assuring that each second- hand BMW 3 Series auto within the Leeds area has an equal chance of being included in the analysis. This approach enhances the representativeness of the sample and minimizes selection bias.

5. Representativeness of the Sample:

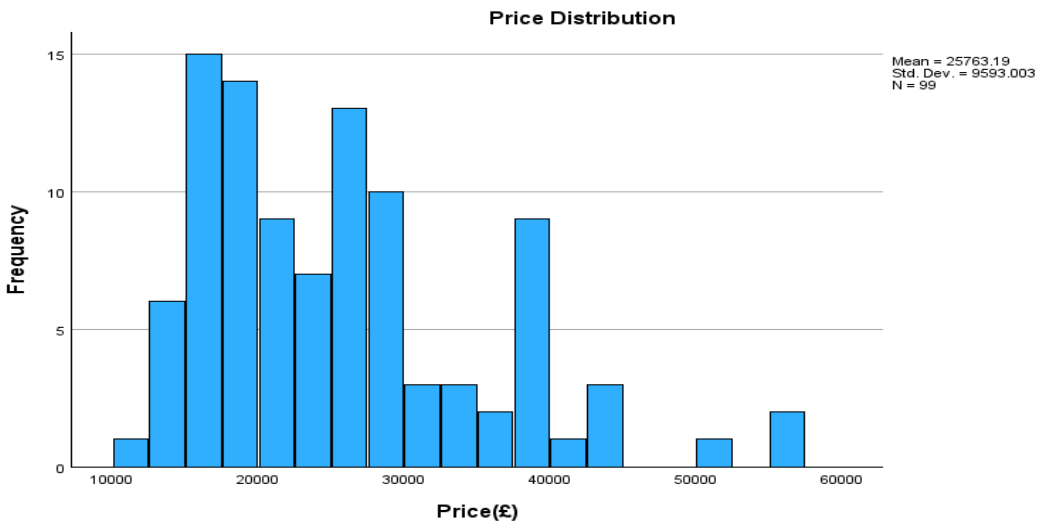
The sample of 300 alternate- hand BMW 3 Series buses is supposed representative of the larger population within Leeds. This is because the random sampling system ensures a different and fair selection, getting a broad range of cars available in the area. The focus on a specific model and area adds in-depth particularity to the analysis impacting pricing within this particular area.

6. Justification for the 5-Year Range:

The decision to limit the sample to cars within the 2018 to 2022 range is predicated in the desire to concentrate on fairly recent market trends. This ensures that the analysis reflects current consumer preferences and aligns with the manager's ambition of understanding the determinants of second- hand car prices in contemporary requests.

A. Visualization Of the Data:

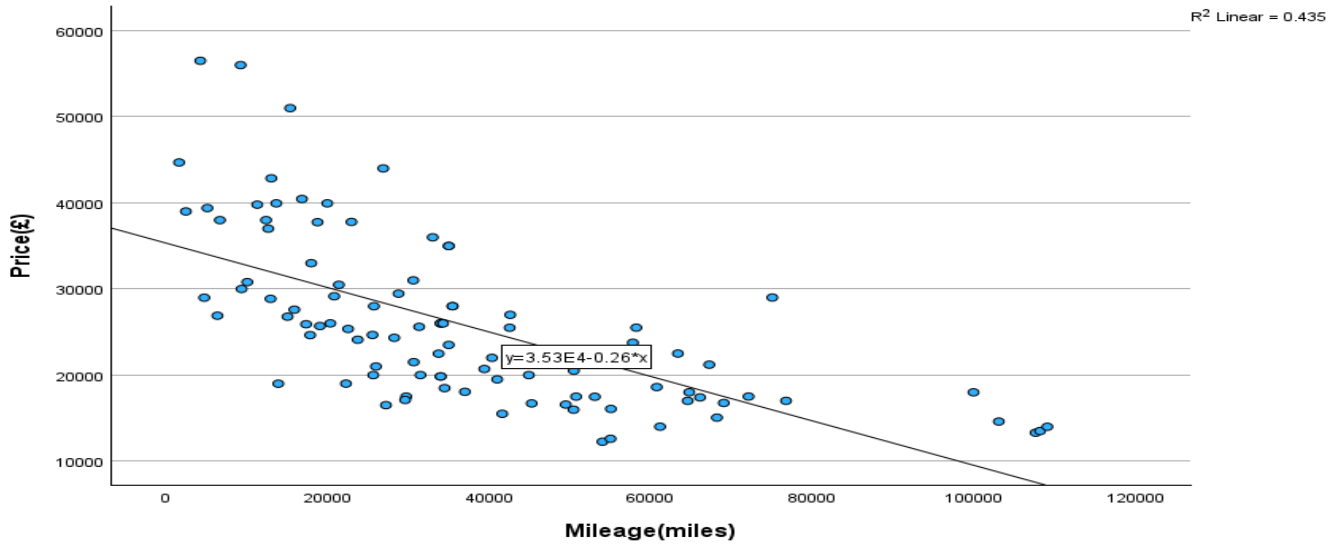
1. Price Distribution:



The bar chart gives a visual representation of the distribution of second-hand car prices based on frequency. Each bar corresponds to a specific price range, showcasing how frequently cars fall within different price brackets in the market.

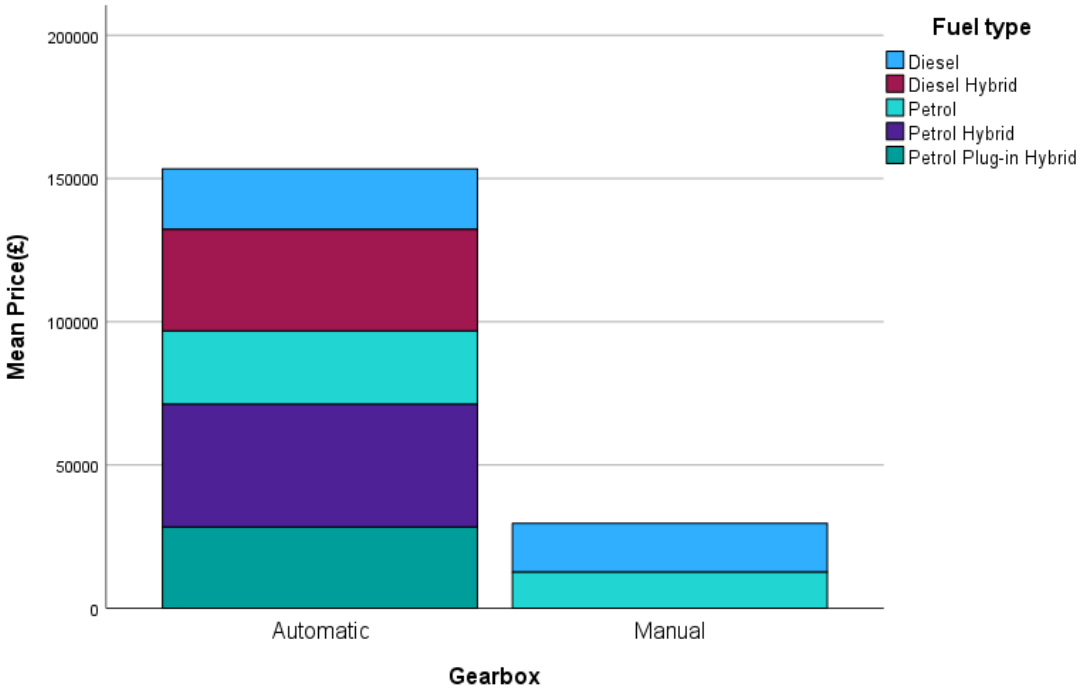
2. Scatter plot of Mileage vs Prices:

The scatter plot visually depicts the relationship between the mileage and price of second-hand BMW 3 Series cars.



As mileage increases, there is a discernible trend indicating that the prices of second-hand cars tend to decrease. One factor is that cars with higher mileage are more likely to have repairs and maintenance issues.

3. Stacked bar chart shows the mean price of a car by fuel type and gearbox.



The most expensive fuel type is diesel, followed by diesel hybrid. The least expensive fuel type is petrol manual.Automatic gearboxes are more expensive than manual gearboxes for all fuel types.

B. Descriptive Statistics for Car Data:

Considering all the variables:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Price(£)	99	12250	56500	25763.19	9593.003	92025712.177
Registration	99	2018	2023	2019.89	1.609	2.590
Engine()	99	1.5	3.0	2.172	.3987	.159
Mileage(miles)	99	1663	109000	37034.42	24465.585	598564865.43
Valid N (listwise)	99					

The dataset shows a diverse range of second-hand cars with varying prices, registration years, engine sizes, and mileages.The standard deviations for price and mileage suggest notable variability, indicating a wide range of pricing and mileage across the dataset.Registration years are relatively close together, with a mean year of 2019.89, indicating a concentration of cars from recent years.Engine sizes are moderately dispersed around the mean, with a relatively low standard deviation.

Considering only engine capacity and mileage:

Report

Price(£)

Engine()	Mean	N	Std. Deviation	Sum	Minimum	Maximum	Range	Median	Variance
1.5	14547.00	2	2761.959	29094	12594	16500	3906	14547.00	7628418.000
2.0	24165.14	79	7709.683	1909046	12250	44690	32440	22790.00	59439214.557
3.0	34023.11	18	12586.655	612416	16060	56500	40440	34989.00	158423878.22
Total	25763.19	99	9593.003	2550556	12250	56500	44250	24322.00	92025712.177

Conclusion:

- Generally, cars with larger engines incline to have higher average prices and more price variability. The total spending is highest for cars with a 2.0 engine.
- The average price for all cars in the dataset is £25,763. This gives a idea of the typical cost that people are willing to pay for a car.
- Car prices in the dataset vary widely, ranging from £12,250 to £56,500. This shows that there are both affordable and high-end options.
- The variance in prices, with a total spread of £92,025,712, indicates that there is a significant diversity in car prices. Some cars are priced quite differently from the average.

- The median price, which is £24,322, gives us the middle point in the distribution. It's less affected by extreme values, providing a strong measure of central tendency.
- The standard deviation of £9,593.00 quantifies the amount of variation or dispersion in car prices. A higher standard deviation suggests greater variability.
- With a total of 99 observations (cars), this dataset provides a considerable sample size for analysis.

C. Confidence Interval:

Descriptives

			Statistic	Std. Error
Price(£)	Mean		25763.19	964.133
	95% Confidence Interval for Mean	Lower Bound	23849.90	
		Upper Bound	27676.48	
	5% Trimmed Mean		25091.40	
	Median		24322.00	
	Variance		92025712.177	
	Std. Deviation		9593.003	
	Minimum		12250	
	Maximum		56500	
	Range		44250	
	Interquartile Range		11995	
	Skewness		1.053	.243
	Kurtosis		.915	.481

From the table above :

- Mean Price (£): £25,763.19
- Confidence Interval: 95%
- Lower Bound: £23,849.90
- Upper Bound: £27,676.48

- **Explanation:**

1. Interpretation of the Confidence Interval:
- The 95% confidence interval shows a range within which we can practically expect the true average price of second-hand cars in the population.
- Here 95% confident that the true mean price is between £23,849.90 and £27,676.48.
- The 95% confidence level is a common standard in statistical analysis. It means that if we were to take many samples from the population and compute a 95% confidence interval for each sample, we would expect approximately 95% of those intervals to contain the true population mean. (Mcleod, Aug 2023)

-The 95% confidence interval for the average second-hand car price suggests that, based on the sample data, we are reasonably confident that the true average price in the population falls within the specified range. This information is valuable for decision-making and understanding the precision of the estimate.

D. Hypothesis Testing:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Price(£)	99	25763.19	9593.003	964.133
Price	309	26422.31	9290.151	528.498

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Price(£)	Cohen's d	9593.003	2.686	2.260	3.108
	Hedges' correction	9667.208	2.665	2.243	3.084
Price	Cohen's d	9290.151	2.844	2.593	3.094
	Hedges' correction	9312.850	2.837	2.587	3.087

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

One-Sample Test

Test Value = 0

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Price(£)	26.722	98	<.001	<.001	25763.192	23849.90	27676.48
Price	49.995	308	<.001	<.001	26422.311	25382.39	27462.23

Hypothesis Testing:

Null Hypothesis (H0): The average price of our chosen car model and the average price in the UK is same .

Alternative Hypothesis (H1): The average price of our chosen car model and the average price in the UK is different .

One-Sample T-Test Results:

For Our Chosen Car Model:

- Significance (p-value): <.001 (Highly significant)
- Mean Difference: £25,763.192

For UK Average Car Price:

- Significance (p-value): <.001 (Highly significant)
- Mean Difference: £26,422.311

Interpretation:

The p-values are significant for both the sample car model and the population car model. This indicates that there is a noteworthy difference in average prices.

The mean difference provides the magnitude of this difference.

Conclusion:

Here the p-significantt value is less than the critical value(0.05) so we reject the null hypothesis. Therefore , the avreage price of car from sample is different than the average price of car from the population.

- This analysis is important for setting up the prices of cars.
- Understanding the difference in average prices is crucial for strategic pricing decisions and market positioning.

E. Correlation Analysis :

The aim of this correlation analysis is to identify the relationships between the price of second-hand cars and various key characteristics, including registration year, mileage, and engine capacity. It dpicts the significance of the factors that can affect the car prices

Descriptive Statistics:

Descriptive Statistics

	Mean	Std. Deviation	N
Price(£)	25763.19	9593.003	99
Registration	2203.73	1828.978	99
Mileage(miles)	37034.42	24465.585	99
Engine()	2.172	.3987	99

Correlation Matrix:

Coreelation for Price(£) and Registration Year:

- Correlation Coefficient: -0.062

Correlations

		Price(£)	Registration	Mileage(miles)	Engine()
Price(£)	Pearson Correlation	1	-.062	-.659**	.427**
	Sig. (2-tailed)		.542	<.001	<.001
	N	99	99	99	99
Registration	Pearson Correlation	-.062	1	-.013	-.044
	Sig. (2-tailed)	.542		.898	.668
	N	99	99	99	99
Mileage(miles)	Pearson Correlation	-.659**	-.013	1	-.085
	Sig. (2-tailed)	<.001	.898		.400
	N	99	99	99	99
Engine()	Pearson Correlation	.427**	-.044	-.085	1
	Sig. (2-tailed)	<.001	.668	.400	
	N	99	99	99	99

**. Correlation is significant at the 0.01 level (2-tailed).

-**Conclusion:** There is a very weak negative correlation between the price and registration year. It shows that older cars may tend to have slightly lower prices, but the relationship is not statistically significant.

-<u>Correlation for Price(£) and Mileage</u>:

- Correlation Coefficient: -0.659

- Significance (p-value): <0.001

- Conclusion: There is negative coreelation between price and mileage.It shows that cars with higher mileage generally have lower prices.

<u>Correlation for Price(£) and Engine Size:</u>

- Correlation Coefficient: 0.427

- Significance (p-value): <0.001

- Conclusion: A significant moderate positive correlation is observed between the price and engine size. Cars with larger engines tend to have higher prices in the second-hand market.

Overall conclusion:

The significant negative correlation between price and mileage has a substantial influence on the pricing of second-hand cars. A significant positive correlation between price and engine size indicates that cars with large engine capacity have the higher price.

It reflects that there is a preference for more powerful vehicles in the market.

Correlation does not imply causation, and other unexplored factors may influence pricing.

- The findings are based on the available dataset, and variations may exist in the broader market..

F. **Regression analysis** :

The regression analysis hels to find the relationships between the price of second-hand cars and key characteristics, including registration year, mileage, and engine power. This investigation is crucial for understanding how these variables collectively impact the pricing of second-hand cars in our dataset.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.759[a]	.575	.562	6348.156

a. Predictors: (Constant), Registration, Mileage(miles), Engine()

b. Dependent Variable: Price(£)

Regression Analysis Report

Introduction:

The regression analysis aimed to uncover the relationships between the price of second-hand cars and key characteristics, including registration year, mileage, and engine specifications. This analysis is pivotal for understanding how these variables collectively influence the pricing of second-hand cars in our dataset.

Model Summary:

R-Square: 0.575 -Indicates that the model explains 57.5% of the variance in car prices.

Adjusted R-Square: 0.562- Adjusts R-Square for the number of predictors in the model.

Std. Error of the Estimate: £6348.156- Represents the standard deviation of the residuals, providing a measure of how well the model predicts the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	16151.572	3889.193		4.153	<.001
	Mileage(miles)	-.246	.026	-.628	-9.362	<.001
	Engine()	8914.084	1615.778	.371	5.517	<.001
	Registration	-.284	.351	-.054	-.808	.421

a. Dependent Variable: Price(£)

Coefficients:

Constant (Intercept): Represents the expected price when all other predictors are zero.Estimate: £16,151.572 ($p < 0.001$)

Mileage: Each additional mile is associated with a decrease in price.Estimate: -£0.246 per mile ($p < 0.001$)

Engine: Shows a positive relationship between engine size and price indicates that cars with larger engines tend to have higher prices.

Estimate: £8,914.084 per unit increase ($p < 0.001$)

Registration: No statistically significant relationship with price suggesting that, in this dataset, registration year does not significantly impact price.

Estimate: -£0.284 per year ($p = 0.421$)

The statistical model derived from the coefficients can be expressed as follows:

$$\text{Price(£)} = 16151.572 - 0.246 \times \text{Mileage(miles)} + 8914.084 \times \text{Engine()} - 0.284 \times \text{Registration}$$

Here,

Price(£) - predicted car price.

Mileage(miles) mileage of the car in miles.

Engine()engine scapacity of the car.

Registration- Registration is a binary variable indicating whether the car is registered or not.

For example –

To predict the price of a car with the following features:

Mileage(miles): 30,000 miles

Engine(): 2.0

Registration: 1 (assuming it's a registered car)

Substitute these values into the model:

Price(£)= 16151.572−0.246×30000+8914.084×2.0−0.284×1Price(£)= 16151.572 − 0.246 × 30000 + 8914.084×2.0− 0.284×1

Price(£)=16151.572−7380+17828.168−0.284

=26600.456−0.284Price(£)

=26600.456−0.284

=26300.172

This is the predicted price of chosen car model.

G. Residual Analysis:

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	6558.03	41264.29	25763.19	7277.381	99
Residual	-15352.707	15962.901	.000	6250.235	99
Std. Predicted Value	-2.639	2.130	.000	1.000	99
Std. Residual	-2.418	2.515	.000	.985	99

a. Dependent Variable: Price(£)

Residual analysis:

Range: Residuals, which represent the differences between predicted and actual prices, vary from -£15,352.71 to £15,962.90.

On average, the residuals are close to zero (0.000), indicating that, overall, the predictions are accurate.

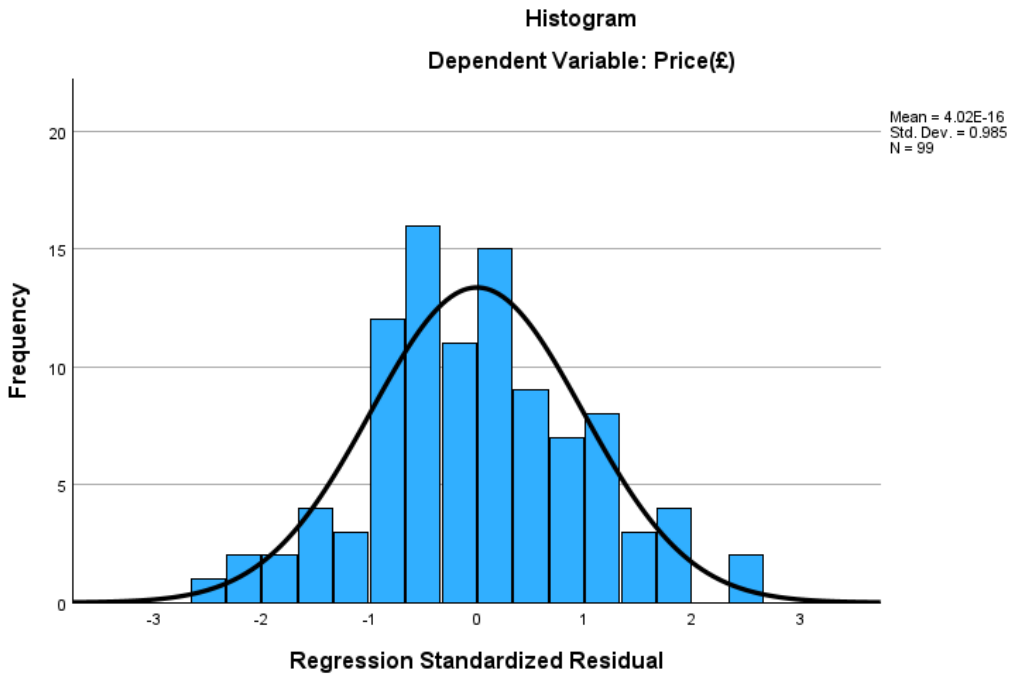
Standardized Residual:

Range: Standardized residuals, which show the size of prediction errors in standard deviation units, range from -2.418 to 2.515.

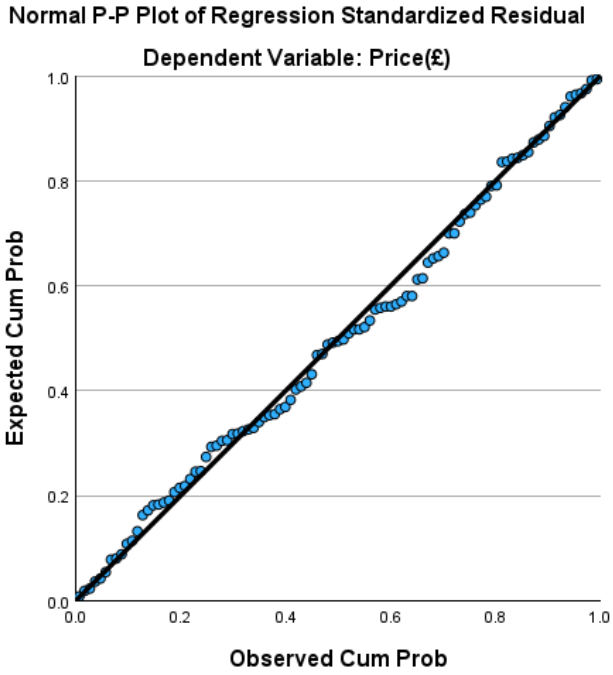
Average: On average, standardized residuals are close to zero (0.000).

Conclusion:

- The predicted price shows variability in car prices.
- Residuals close to zero for each shows that predictions are accurate.
- These statistics provide a detailed look at how well the prediction model is performing and the variability in the accuracy of predicted car prices.
- In conclusion, these statistics give us a detailed view of our model fit.
- The most parsimonious model is the model that has all of the coefficients of the independent variables significantly different from zero. Our model has all the independent variables different than zero so our chosen model is the most



parsimonious model.



H. Assumptions of Linearity :

The assupmtions of linearity are as follows

1. Linearity:

- Assumption:The relationship between the independent variable(s) and the dependent variable is linear.

Interpretation: The regression line is linear and it accurately shows the true relationship between variables.

2. Independence of Residuals:

- Assumption:Residuals (the differences between observed and predicted values) should be independent of each other

- Interpretation: Residuals close to zero for each shows that predictions are accurate.

3. Normality of Residuals:

- Assumption: The residuals should be approximately normally distributed.

- Interpretation: Residuals being "approximately normally distributed," there should be a pattern similar to a bell-shaped curve. Also, here mean residual is £0.000, suggesting that, on average, the model doesn't consistently overestimate or underestimate car prices.

4. No Perfect Multicollinearity:

- Assumption:There should be no perfect linear relationship among the independent variables (no perfect multicollinearity).

- Interpretation: Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to isolate their individual effects. Here there is significant relation between variables. Price and engine size.

I. **Justification**:

The regression and residual analysis is performed to evaluate the predicted second-hand car prices. The model is statistically significant, providing insights into the impact of each variable on the car prices.The analysis forms a solid foundation for informed decision-making and future research in the automotive market.

Reference

Mcleod, Saul (PhD). *Confidence Intervals Explained: Examples, Formula & Interpretation*. 11 Aug. 2023, www.simplypsychology.org/confidence-interval.html.