Business report on visualisation and statistical analysis of a real-world dataset

Academic Year 2021/22

MODULE CODE: BNM854J

MODULE NAME: Descriptive Analytics

MODULE LEADER: Leonaidas Anastasakis

By

Candidate Number: 670651

Table of Contents

EXECUTIVE SUMMARY

The report provides an analysis of the second hand car market in the B17 area and it evaluates the average car price in the UK from the sample data. It provides the perception for the factors to keep in mind to buy second hand car from the market. The Statistical methods and data visualization has been used in the report to figure out graphs from the sample and figures which affect the market. After analysing we reported that price of the car decreases if number of miles driven by car increases. Moreover, most of the used cars on sale are of diesel fuel type. But, there is a much difference between the average price of the car which we got from the Statistics and the average price in the UK. It has been observed from the figures that 60% of the car price depend upon the independent variable such as age of car, number of miles driven, engine capacity in L.

CAR MARKET VISUALIZATION

1. INTRODUCTION

1.1 DESCRIPTION OF THE PROBLEM:

The goal of the report is to provide statistical analysis on the UK's second hand car market for the postcode B17. From this report, the manager of the real estate agency can use this report to determine which factors are most important to estimate the market value of a car by knowing its mileage, engine capacity, and vehicle type.

1.2 SOURCE OF DATA:

The data for the analysis was gathered from www.autotrader.co.uk, one of the most reputable real estate websites in the United Kingdom. The data for the specific area is a real data set that is easily accessible on the website and suitable for analysis.

1.3 SAMPLING METHOD:

The method which is used for sampling is **Simple random**. Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample.

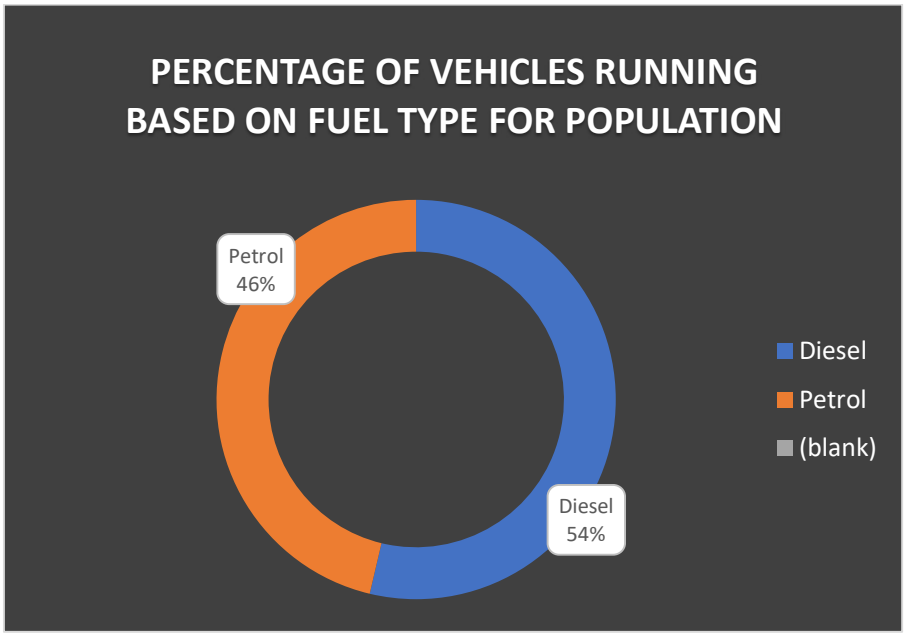
Steps used for Sampling Method:

a. We have used Excel to perform simple random sample to extract a sample data from large population.
b. Prepare a list of a population entry into the column and each entry will be assigned with a specified number.
c. Next column is used for random number to generate random number using rand() formula for all the population.
d. Decide on the sample size and we can either choose sample from top, bottom or centre and in this way sample is selected through randomly generated numbers in Excel.
e. There are several reasons to choose simple random sampling such as:

i. One of the great advantages of simple random sampling method is that it does not need to have a prior knowledge of the subject.
ii. Easy to access sampling error.
iii. This is suitable for data analysis which includes the use of inferential statistics.
iv. Simple random sampling method is easy to use.
v. It is unbiased approach to gather sample data from large population.

2. <u>VISUALIZATION:</u>

<u>Graph1:</u>



Row Labels	Count of Fuel_type
Diesel	179
Petrol	154
(blank)	
Grand Total	**333**

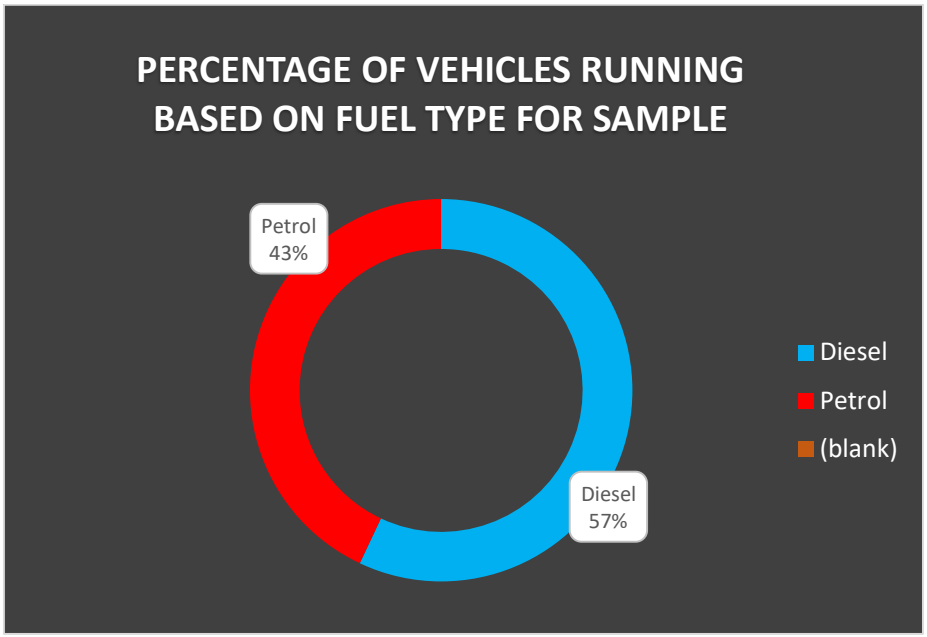
Row Labels	Count of Fuel_type
Diesel	57
Petrol	43
(blank)	
Grand Total	**100**

The pie graph depicts the most common fuel type car owned by the general public and which are on sale as well. According to the graph, the majority of vehicles which are on sale are diesel vehicles over petrol vehicles. Petrol accounts for 46% percent of the total population, while diesel vehicles account for only 54%.

The same pattern can be seen in the sample. This implies that the sample may be a good representation of the population. Its adequate to use pie chart to visualize the structure of market based on fuel type between two sets of data.

Proportionality:

According to the principle, the representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented. So, no concentric doughnuts! Moreover, According to the rule graph should be data driven. It should show variation that is in the data and not in design variation.

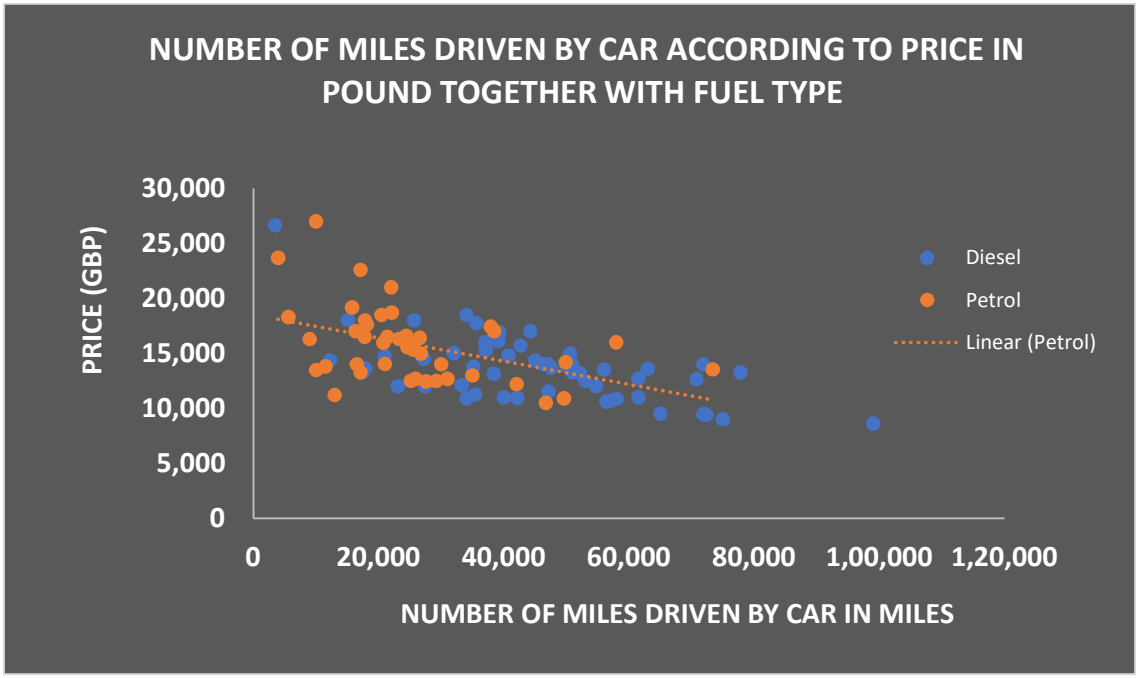


Graph2:

The principle suggests that write out explanations of the data on the graph itself. All important events in the data should be label carefully and axis should start from 0!

The scatter graph shows relation between dependent variable and one or more independent variable.

In this graph it shows negative linear relationship between price of car and number of miles driven by car. We can analyse as the number of miles increases the price for the car decreases. To Compare we have analysed together with the fuel type, petrol as well as diesel. For the the fuel type if number of miles are increased price comes down.

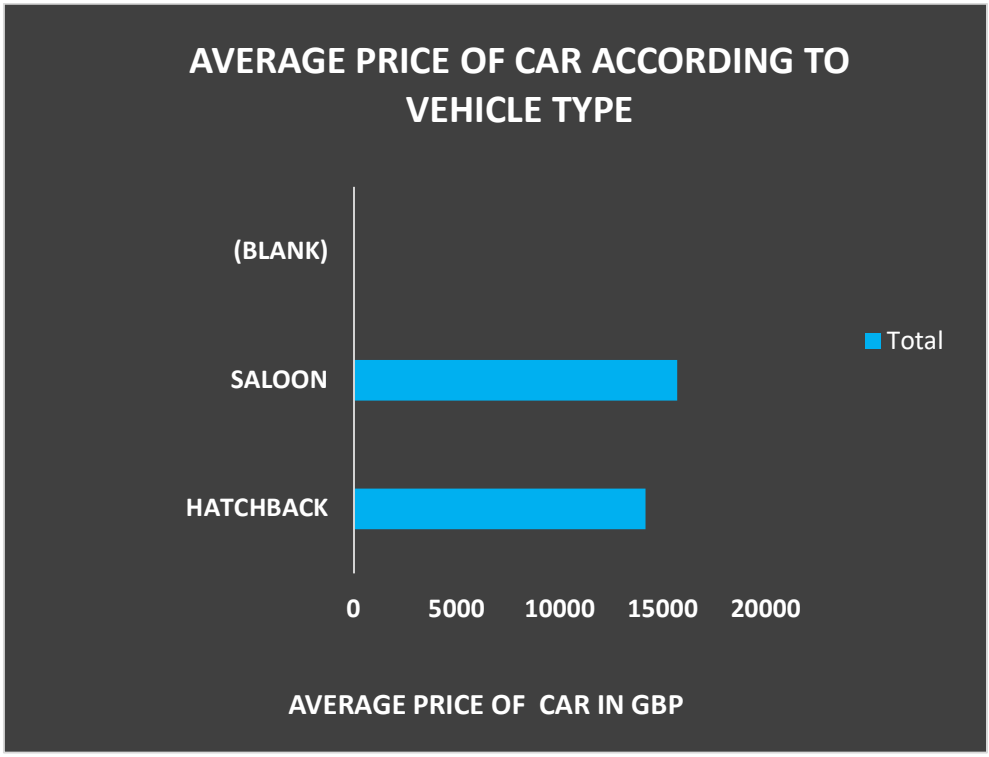


Graph3:

The bar graph represents average price of car according to type of vehicle In B17 postcode.

According to graph, we can interpret that most of the people prefer premium cars rather than smaller sized vehicle. The average price of premium saloon size car is higher than compared to small sized vehicles. The average price of saloon in market is £ 15698.82609 whereas the average price of Hatchback vehicle is £ 14148.22078. To illustrate the categorical data with average price, bar chart is very helpful.

Row Labels	Average of Price in £
Hatchback	14148.22078
Saloon	15698.82609
(blank)	
Grand Total	**14504.86**

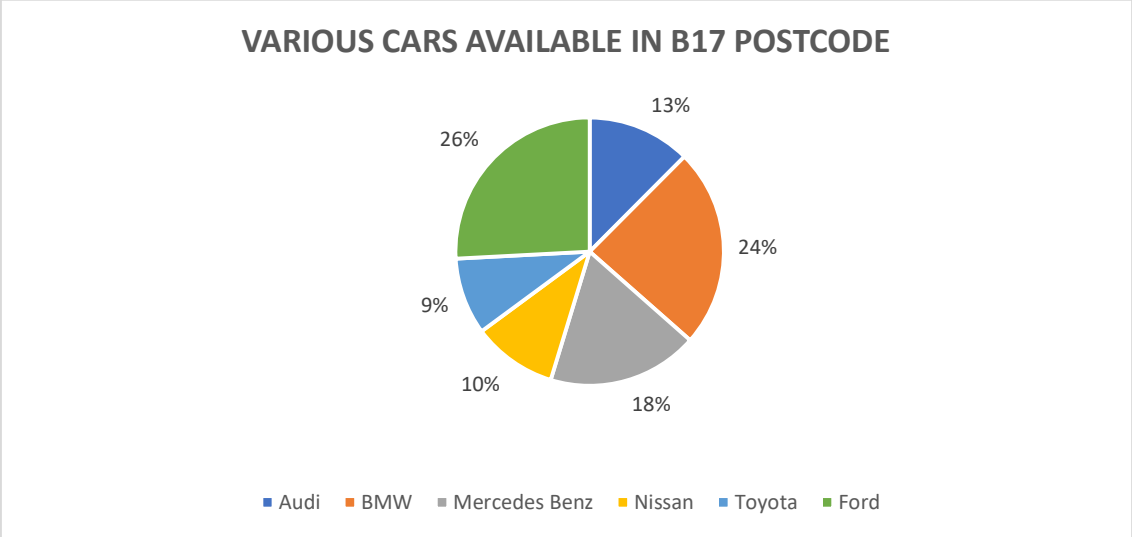


The graph should be clear enough with all the data labels and should display all the trends. So, 3D and perspective effects should be avoided.

Graph4:

Percentage of various types of car available in Harborne B17 postcode.

Car_type	Number_of_car_in market
Audi	1912
BMW	3684
Mercedes Benz	2788
Nissan	1563
Toyota	1420
Ford	3962



3. DESCRIPTIVE STATISTICS TABLE.

The Appendix Table1 displays descriptive statistics for the dependent variable, "price," as well as all the other independent variables, such as the age of the car, the number of miles driven, and the engine capacity of the car in litres. Using these observations, we can derive information for our analysis.

3.1. Mean: The arithmetic mean is calculated by adding up all the observations and dividing by the number of observations. As mentioned in the given table the average price for the car in the market in the postcode B17 is £ 14504.86, an average age of cars is 4.02, average number of miles driven is 36833.26 miles and the arithmetic mean for engine capacity is 1.606. Moreover, we can use the arithmetic mean table to analyse the car market.

3.2. Median: In Statistics, the median is calculated as the middle value of the sorted list. Basic functionality of median is to describe dataset by comparing to the mean. As shown in (APPENDIX TABLE 1), the median for the price of car is £ 14000 which is very close to mean value and from this we can conclude that data distribution is quite evenly distributed.

3.3. **Mode**: It is defined as the observation (or observations) that occur with the greatest frequency in the dataset. As we can see that mode for the price of car is £ 14000 which indicate that most of the cars have same market value. Similarly, from the descriptive statistics table we can get an idea that most of the cars in the market have age 5 and engine capacity of 2L.

3.4. **Standard Deviation**: The standard deviation measures how evenly distributed the sample data is within a given range of values. Our standard deviation for price of car is £ 3359.436298, according to our analysis. This means that this value represents the average deviation from the mean price value in our data. We can draw the same conclusions from the table above for our independent variables.

4. 95% CONFIDENCE INTERVAL FOR THE AVERAGE CAR PRICE

4.1 CONFIDENCE INTERVAL

A **confidence interval** (**CI**) is a type of estimate computed from the statistics of the observed data. The probability that the population parameter is within the interval, is usually expressed as a percentage, which is called the confidence level.

In our analysis we have calculated the confidence interval with 95% confidence level and used z-test for the calculation. The z-test is the statistical test which gives us the confidence level value. We have to used z-test because our sample size is more than 30 and we have a sample of 100 cars.

From the calculation of confidence interval, we got the range of 13838.2750 to 15171.4450. This means that we can be 95% certain that the average car price for all types in Harborne B17 area lies between the calculated range. However, if the confidence level is 95%, the significance level would be 5%. It states that there is a still chance that population mean will not fall under this confidence interval range.

5. COMPARISON OF AVERAGE CAR PRICE WITH THE AVERAGE PRICE IN THE UK

For requirement to know whether the average price of the different type of cars in Harborne, Birmingham B17 area sample is in line with the average car price in the UK (or your region/county). We have carried out hypothesis testing.

In statistics, hypothesis testing is the process by which an analyst tests an assumption about a population parameter. The methodology used by the analyst is determined by the nature of the data and the purpose of the analysis. Using sample data, hypothesis is used to assess the plausibility of a hypothesis.

5.1 Important terms to understand:

a) **Null hypothesis**: It is a theory that someone wants to test. It is assumed to be true until proved otherwise based upon your sample evidence.
b) **Alternative hypothesis**: It states that a population parameter is smaller, greater, or different than the hypothesized and it is not equal to hypothesized value.
c) **Critical value**: It is a point on the test distribution that is compared to the test statistic to determine whether to reject the null hypothesis.

5.2 ONE SAMPLE T-TEST FOR AVERAGE PRICE: To conduct a test we have to specify null alternative hypothesis and specify the level of significance which is 5% in our case. After applying the test statistic, we get the value and we need to compare that value with our critical value. If the calculated value is greater than the critical value, we reject the null hypothesis and if it is less than the critical value, we accept the null hypothesis. From the hypothesis test we have concluded that we accept the null hypothesis as we got the calculated value of 0.000(see Appendix table 2) which is less than our 0.05 which means that there is sufficient evidence to suggest that average price of car in a specific area.

Average price of car £ 21355.833 which is greater than our calculated sample mean price of 14504.86. We can have an idea from the data that there is quite difference between our sample mean price and the average price of car in the UK.

5.3 Source of Data: The data for the average price of car in the UK has been obtained from the website https://www.carsite.co.uk/.

6. CORRELATION ANALYSIS

A correlation is an approach to describe the relationship between two variables. Any value in the correlation table greater than 0.7 or less than - 0.7 are significant. Any correlation between range 0.7 and -0.7 is statistically insignificant.

Correlation analysis is required to determine how our independent variable such as age of car, number of miles driven, and engine capacity are affected from dependent variable "price". According to the correlation matrix (see Appendix Table 3), we have figured out that there is moderately correlation between age of cars and price as the value we got from the correlation analysis is -0.661 which is almost equal to value of 0.7. We can analyse on the basis of correlation matrix table about the correlation among all independent variable engine capacity in L and car price have really weak correlation among them as the values are less than the 0.7.

7. REGRESSION ANALYSIS AND THE PARSIMONIOUS MODEL

7.1 REGRESSION ANALYSIS

The software Package Used for Regression Analysis is **SPSS**. In statistics, regression analysis is a method to identify the relationships between a dependent variable and one or more independent variables. It allows you to analyse the strength of variables and which factors are important for visualisation.

Regression analysis is required to derived the most parsimonious model. The parsimonious model is used to check whether all independent variables are significant. To do that we can either look at t value or sig value in of the Coefficients in our model. We need to compare the sig values of this column to 0.05 values to check the significance. If the sig column value is higher than the 0.05 then the variable is insignificant and if it is lower than 0.05 it is significant, or the t value should be in range $-1.96 < t < +1.96$.

Based on our findings, we concluded that our independent variables, such as car age, number of miles driven, and engine capacity, had significance values lower than the 0.05 significance value explained above. This means that all these independent variables are significant and have a significant effect on the price of the car on which we rely.

7.2 STEPS TO DERIVE THE MOST PARSIMONIOUS MODEL

To derive the most parsimonious model, we will remove the most insignificant variable from the model whose value is greater than 0.05(refer Appendix table 4) and we need to run regression model again. Every time we run the model we need to check if the coefficients and check the t or sig value of independent variables. When the sig values for all the independent variables in coefficient table are lower than 0.05, then we get the most parsimonious model (see appendix Table 5) with the independent variable. We got our the most parsimonious model in the second time because variable such as 'age' with the significance value of 0.00, 'number of miles' with the significance value of 0.000 and 'engine capacity' with the significance value of 0.00 respectively which is less than 0.05 significance value and variable which have higher value than 0.05 such as fuel type and vehicle type, we have removed those variables from model to get most parsimonious model.

7.3 DESCRIPTIVE STATISTICS:

From (Appendix table 7) we can derive mean, Standard deviation, and N for the sample data for dependent and independent variables.

7.4 MODEL SUMMARY: According to (Appendix table 8) , it describes model summary for which the 60.0 % variation is explained by this model. We need to look at other factor as well because 40% is not explained by independent variables that has a linear effect on dependent variable.

7.5 RELIABILITY OF THE INDEPENDENT VARIABLE:

Result from ANOVA (appendix Table 9) analysis shows that all the independent variables are reliable to explain the price of second hand car in the market.

7.6 SIGNIFICANCE OF THE INDEPENDENT VARIABLE:

p- value for all the independent variables are less than 0.05. This conclude that price is affected by all the other independent variables.

8. RESIDUAL ANALYSIS FOR THE MOST PARSIMONIOUS MODEL

8.1 RESIDUAL ANALYSIS:

We can now test whether we have the most parsimonious model or not after we have obtained the most parsimonious model. It is critical to first assess the model's adequacy before assessing its goodness. **Multicollinearity** is the occurrence of high inter correlations among two or more independent variables in a multiple regression model issues here because our independent variables are not correlated with each other, So we will be only focusing on examining following four assumptions of Residual Diagnostic test to check the model`s adequacy.

8.1.1 **Assumption 1:** Linearity: The assumption of linear regression suggests that the average of residual must be equal to zero or we can say is that it should be scatter around 0.

8.1.2 **Assumption 2:** The error terms (residuals) are independent. If you observe a random scatter diagram then the assumption will hold. If you observe a clear pattern in the residuals in the form e.g. four residuals are above zero, then four below, then four above and so on, the assumption will fail.

8.1.3 **Assumption3**: The standard deviation of the residuals (ε) is the same for all values of the estimated dependent variable. If the distance of the residuals in the graph in relation to zero remains more or less the same as the estimated values of the dependent variable increase, then the assumption holds. If you can see a clear pattern of the residuals increasing or decreasing as the estimated values of the dependent variable increase, then the assumption will fail.

8.1.4 **Assumption4:** The residuals ε is normally distributed. For assumption 4 we see a histogram of the residuals and we investigate whether they are normally distributed. We can also see a normal P-P plot to see whether a large deviation exists between our standardised residuals in relation to the residuals if the data was perfectly distributed.

8.2 ADEQUACY OF THE DERIVED MODEL

From our analysis of standardised residuals (see appendix Graph1) we observe that residuals are randomly scattered over 0. Model hold for all the assumptions; therefore, this model is adequate. The model is adequate for use for further predictions.

9. GOODNESS OF THE MODEL

To check the goodness of the model we see the adjusted R2 value (see appendix Table 7). Adjusted R2 value is 0.600. This means that the model is only 60.0% variation is explained by this model. We need to look at other factor as well because 40% is not explained by independent variables that has a linear effect on dependent variable.

10. DERIVED STATISTICAL MODEL AND JUSTIFICATION OF THE USAGE

The derived model is adequate and is good for further prediction of price as it holds the assumptions of adequacy (appendix Graph1) and R2 value (see appendix Table 7) is quite good. The model can be used for further predictions as the independent variables such as age of car, engine capacity and number of miles can be used as good predictors for analysing the model.

Appendix:

Table1:

Statistics	Price in £	Age	Miles_Driven	Engine_Capacity in L
Mean	14504.86	4.02	36833.26	1.606
Standard Error	335.9436298	0.099473361	1938.658102	0.0326233
Median	14000	4	35059.5	1.6
Mode	14000	5	72320	2
Standard Deviation	3359.436298	0.994733607	19386.58102	0.3262335
Sample Variance	11285812.24	0.989494949	375839523.6	0.1064283
Kurtosis	2.703537275	0.686296532	0.011236769	-0.6740339
Skewness	1.198527885	-0.92023871	0.609387689	-0.3217014
Range	18391	4	95551	1
Minimum	8599	1	3449	1
Maximum	26990	5	99000	2
Sum	1450486	402	3683326	160.6
Count	100	100	100	100

Table 2:

One-Sample Test

Test Value = 21355.833

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Price_of_car	-20.393	99	0.000	-6850.97300	-7517.5580	-6184.3880

Table 3:

Correlations

		Car_price	Age_Car_inyears	Mileage_in_miles	Engine_Capacity_in_L
Pearson Correlation	Car_price	1.000	-0.661	-0.600	-0.092
	Age_Car_inyears	-0.661	1.000	0.478	0.314

	Mileage_in_miles	-0.600	0.478	1.000	0.470
	Engine_Capacity_in_L	-0.092	0.314	0.470	1.000
Sig. (1-tailed)	Car_price		0.000	0.000	0.182
	Age_Car_inyears	0.000		0.000	0.001
	Mileage_in_miles	0.000	0.000		0.000
	Engine_Capacity_in_L	0.182	0.001	0.000	
N	Car_price	100	100	100	100
	Age_Car_inyears	100	100	100	100
	Mileage_in_miles	100	100	100	100
	Engine_Capacity_in_L	100	100	100	100

Table 4:

Coefficient table for all variables.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	18616.589	1531.246		12.158	.000	15576.264	21656.914		
	Age_Car_inyears	-1690.617	249.393	-.501	-6.779	.000	-2185.792	-1195.442	.734	1.362
	Mileage_in_miles	-.087	.014	-.503	-6.401	.000	-.114	-.060	.647	1.545
	Engine_Capacity_in_L	3116.794	743.262	.303	4.193	.000	1641.031	4592.558	.768	1.301
	Fuel_type	713.193	432.485	.106	1.649	.102	-145.516	1571.902	.976	1.025
	Vehicle_type	-104.206	513.364	-.013	-.203	.840	-1123.501	915.090	.958	1.043

a. Dependent Variable: Car_price

Table 5:

Coefficient table for **parsimonious model**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	19701.411	1259.515		15.642	.000	17201.293	22201.530		
	Age_Car_inyears	-1758.011	245.971	-.521	-7.147	.000	-2246.259	-1269.763	.761	1.314
	Mileage_in_miles	-.086	.014	-.494	-6.300	.000	-.112	-.059	.658	1.521
	Engine_Capacity_in_L	3126.625	746.211	.304	4.190	.000	1645.407	4607.842	.769	1.301

a. Dependent Variable: Car_price

Table 6:

Correlation table for parsimonious model

Correlations

		Car_price	Age_Car_inyears	Mileage_in_miles	Engine_Capacity_in_L
Pearson Correlation	Car_price	1.000	-0.661	-0.600	-0.092
	Age_Car_inyears	-0.661	1.000	0.478	0.314
	Mileage_in_miles	-0.600	0.478	1.000	0.470
	Engine_Capacity_in_L	-0.092	0.314	0.470	1.000
Sig. (1-tailed)	Car_price		0.000	0.000	0.182
	Age_Car_inyears	0.000		0.000	0.001
	Mileage_in_miles	0.000	0.000		0.000
	Engine_Capacity_in_L	0.182	0.001	0.000	
N	Car_price	100	100	100	100
	Age_Car_inyears	100	100	100	100
	Mileage_in_miles	100	100	100	100
	Engine_Capacity_in_L	100	100	100	100

Table 7:

Descriptive Statistics

	Mean	Std. Deviation	N
Car_price	14504.8600	3359.43630	100

Age_Car_inyears	4.0200	.99473	100
Mileage_in_miles	36833.2600	19386.58102	100
Engine_Capacity_in_L	1.6060	.32623	100

Table 8:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.783[a]	.613	.600	2123.60089

a. Predictors: (Constant), Engine_Capacity_in_L, Age_Car_inyears, Mileage_in_miles

b. Dependent Variable: Car_price

Table 9:

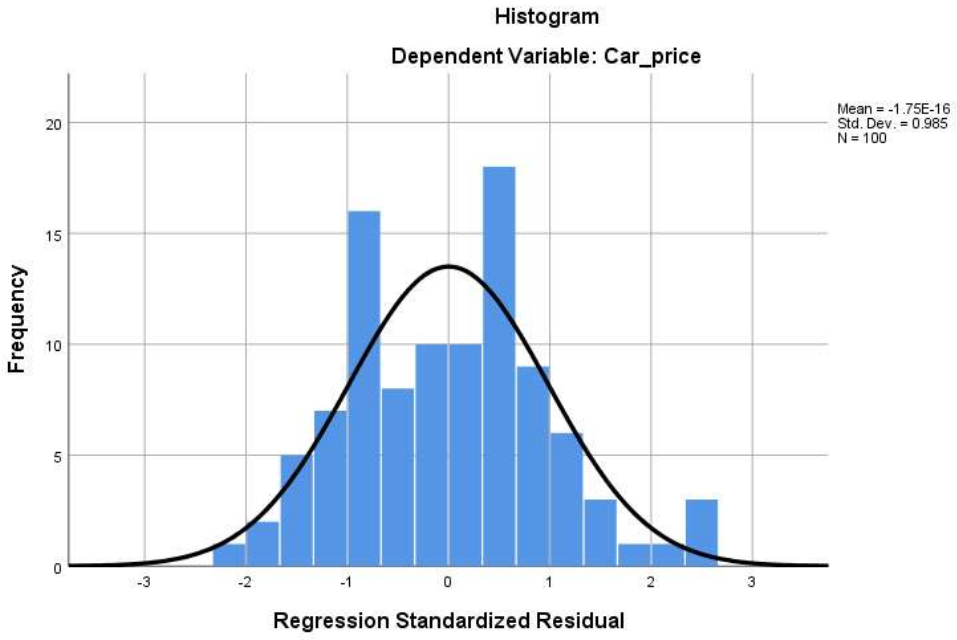
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	684366062.6	3	228122020.9	50.585	.000[b]
	Residual	432929349.4	96	4509680.723		
	Total	1117295412	99			

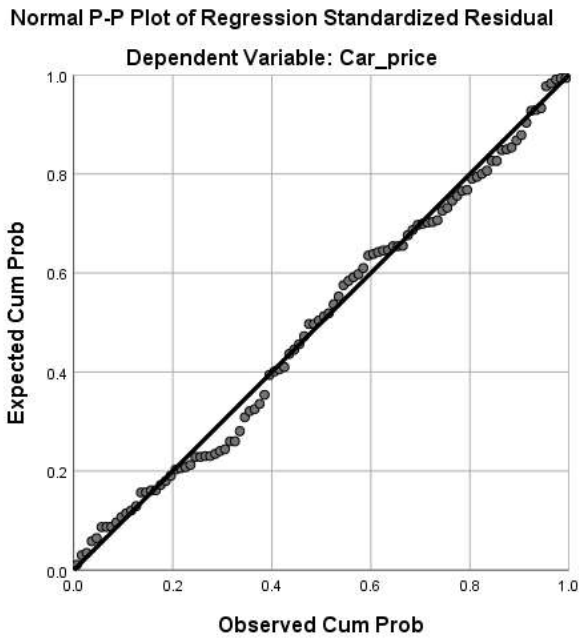
a. Dependent Variable: Car_price

b. Predictors: (Constant), Engine_Capacity_in_L, Age_Car_inyears, Mileage_in_miles

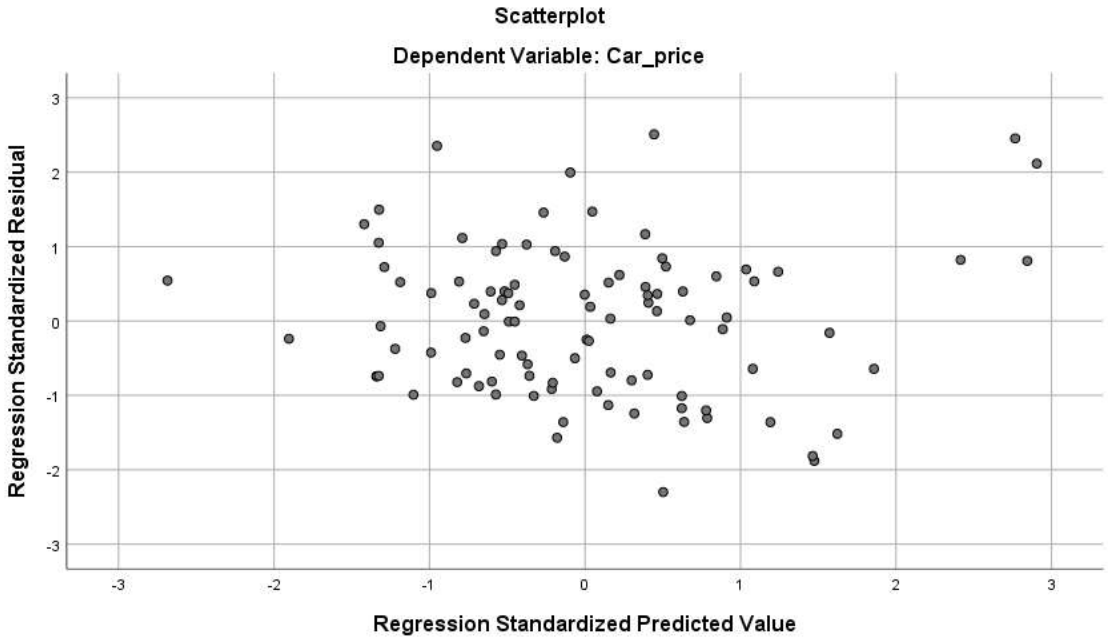
Graph1:



Graph 2:


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Car_price
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

Graph 3:


Scatterplot
Dependent Variable: Car_price
Regression Standardized Residual
Regression Standardized Predicted Value
-3
-2
-1
0
1
2
3