Business report

on

2nd hand car market in Birmingham B29 7RG

Candidate number : 220269920

BNM854 Descriptive Analysis

Contents

Introduction

The information for this sample fleet of vehicles was gathered from the auto marketplace website Auto Trader. The website enables the use of a straightforward random sample, which is an efficient sampling technique to guarantee that automobiles were chosen randomly for the research and provided a representative sample.

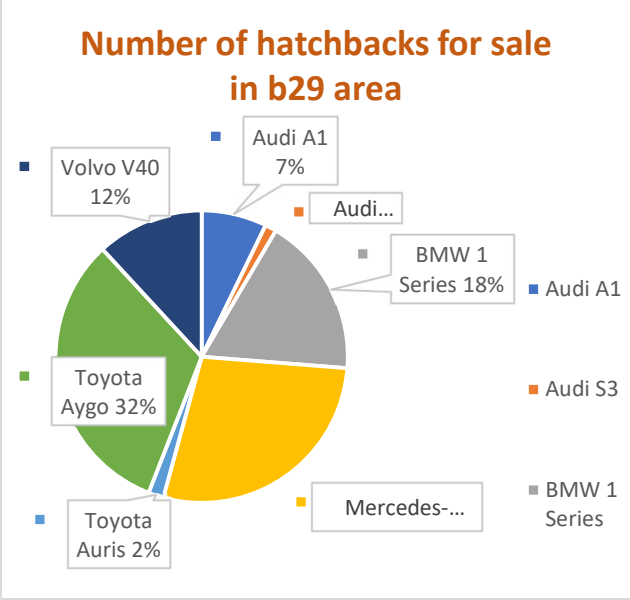
The Mercedes A-class with the zip code B297RG is the vehicle example I chose from the Auto Trader website. Using stratified randomization, I chose 100 A-class vehicles for the sample.

The limitation of data are that many customers also have want to know the average miles per gallon of the car but there was no information on that.

The primary justification for choosing a 5-year range is that, from the standpoint of the customer, only vehicles that are up to five years old are seen as reliable, and those that are older run the risk of malfunctioning or needing maintenance.

Visualization 1

Number of A-class with its competitors in b29 7rg area in the 2nd hand market in last five years.



The above information is in the form of pie chart as The above pie chart shows us the information about the number of cars from different makers in the hatchback category. It is evident that highest number of cars for sale in the 2nd hand market in hatchback category in b29 area was of Toyota Aygo at 32% which was followed by Mercedes A-class 29% which is a luxury hatchback. Also, the number of BMW was also significant at 18% considering it a luxury make as well while on the other hand Audi A1 and Volvo were 19% combined and Audi s3 trailed last at just 1%.

Graphical integrity

The graph has 2 numerical quantities representing and it is a 2-dimension model which is appropriate. the representation of number on the surface of the chart are the same and directly proportional to the numerical quantities represented so the chart follows the rule of proportionality

The labels in the chart are clear and non-distorting and non-ambiguous they determine important data related to the chart and its quantities and therefore the clarity in the chart is appropriate as well.

The chart is data driven, no access or misleading information is provided, instead it provides just the amount of information needed in order to show the user what it is supposed to. Also, it does not manipulate the information by implementing variations in the design.

The graph is contextualised as it gives just the amount of information it is supposed to and does not quote out of context.

Graphical Excellence

The chart gives maximum information with least amount of ink while using least amount of space. Also, there is no repeated information in the chart

Gestalt principles

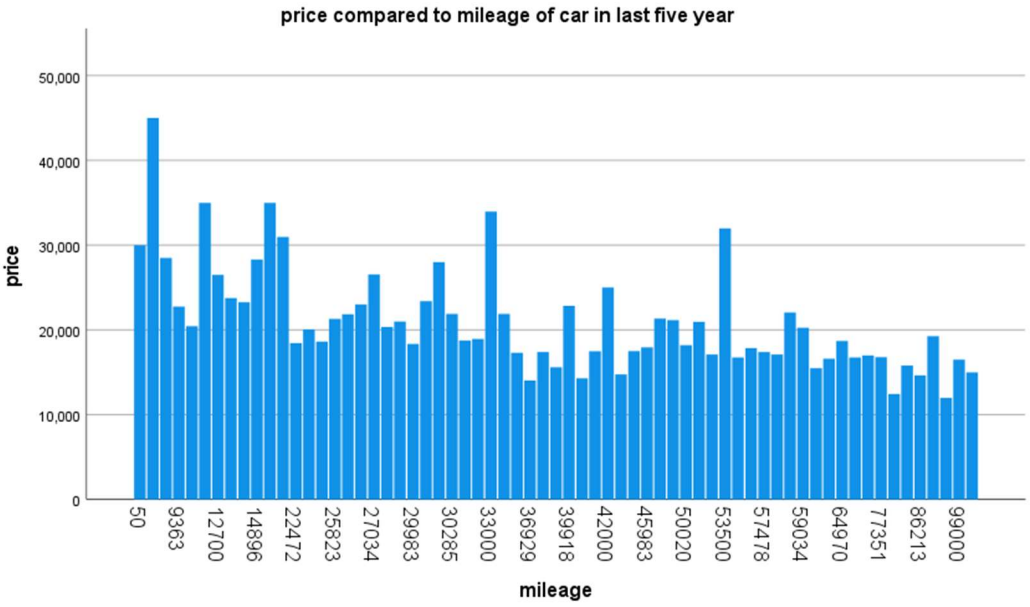
In order to achieve the aim, data for various automobiles of various types is categorised and connections are created, making proper use of proximity.

A circle is shown by each pie slice being constantly slacked together to show continuity.

The graph orientation also follows IBCS rules.

Visualization 2

visualization for price vs mileage



The chart we have used to compare the price and mileage of Mercedes A-class is a column chart because Column charts can show us the data changes over a period of time or for illustrating comparisons among items and we can easily organize the categories along the vertical and horizontal axis.

In the above bar graph, the variation in price of Mercedes A-class with manual and Automatic transmission for the past five years is shown. It is quite evident from the graph that as the Mileage increases the value of that car decreases and therefore the price goes down. For

example, the maximum price from the graph is 45000£ for An A-class hatchback with nearly 50 miles driven but we can get an A-class for as low as 17000£ but the miles driven for that car will be between 80000 to 90000 which is close to overdriven and with this much amount of miles on the dash, the car would need a lots of maintenance and services.

Graphical integrity

The graph uses a two-dimensional model, which is acceptable, and contains the two numerical values "price" and "Mileage." The chart complies with the proportionality rule since the numbers depicted on its surface are same and directly proportionate to the numerical values shown.

There are no labels in the graph as needed information can be clearly seen and therefore the clarity in the chart is appropriate as well.

The chart is data-driven; no access or inaccurate information is given; only the minimum amount of data is given to show the user what is intended. Additionally, it does not alter the content by using different design elements.

The graph is contextualised since it only provides the necessary information and does not quote out of context.

Graphical Excellence

The chart uses the least amount of space. Additionally, the graphic does not contain any redundant or repetitive data, The graph also uses quite some ink but the data shown in the graph is also important and intended data.

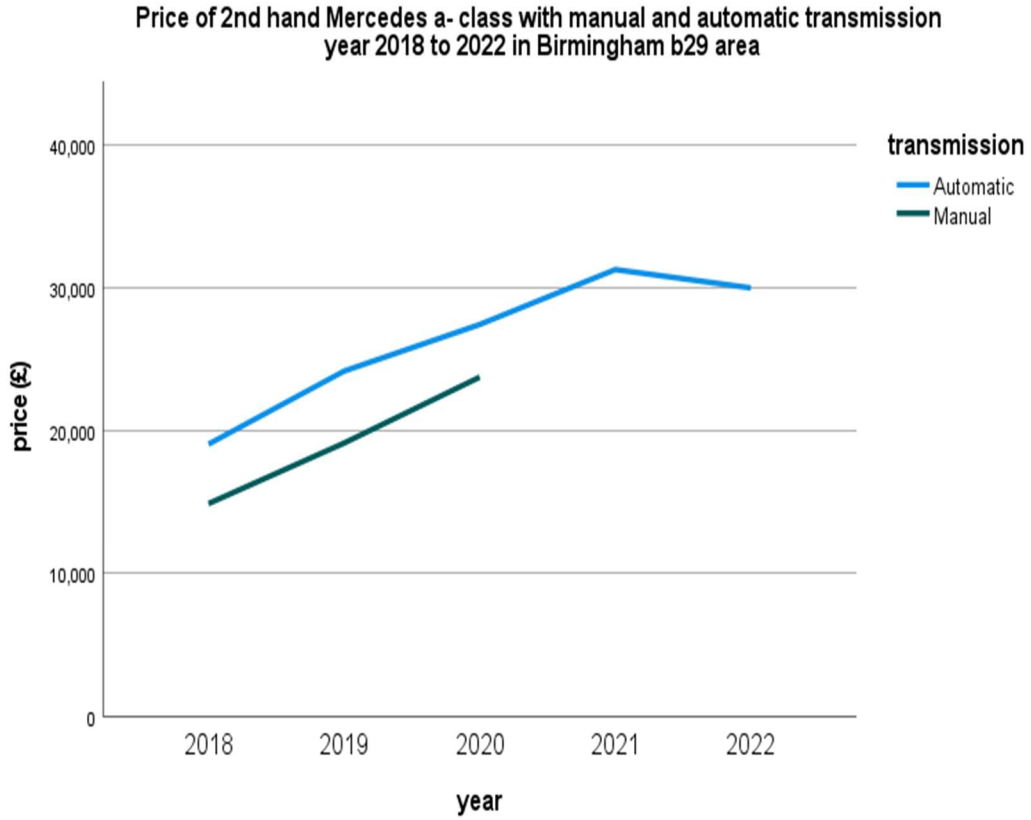
Gestalt principles

This graph adheres to the idea of proximity and similarity since the columns next to each other indicate the closest mileage between them and all of the columns have the same colour signifying the same things (i.e., mileage). Additionally, the graph does not adhere to the idea of connection and continuity since connecting every column to one another would not cause it to behave like a group and the drop in miles with rise in price is not continuous.

The graph orientation also follows IBCS rules.

isualization 3

The increase of price in last five years for manual and automatic transmission



The chart we have used to show the amount of automatic and manual Mercedes A-class is a line graph because line graphs are used to monitor changes over both short and extended time periods. Line graphs are more appropriate for usage when there are smaller changes.

In the above line graph, the variation in price of Mercedes A-class with manual and Automatic transmission for the past five years is shown. It is quite evident from the graph that manual cars in hatchback category were not in sale after 2020 but the price for manual type A-class increased till 2017 to 2020 in the b29 area of Birmingham. While the price of this car has in automatic transmission has seen a steady rise of price from 2017 to 2022.

Graphical integrity

The graph comprises the two numerical values "price" and "year" and employs a two-dimensional model, which is appropriate. Since the numbers portrayed on the chart's surface are identical and directly proportionate to the numerical values displayed, it conforms with the proportionality principle, this graph is also standardised as it uses normal units to represent the monetary value.

Since the necessary information can be viewed clearly without labels, the clarity of the graph is also adequate and therefore the graph follows clarity.

- The chart is data-driven; no erroneous or incomplete information is provided; only the bare minimum of information is required to present the user the desired information. Additionally, the use of different design components does not change the content. The graph is contextualised since it only provides the necessary information and does not quote out of context.

Graphical Excellence

The chart uses the least amount of space and the least amount of ink to convey the most information. On the other side, this graph shows us quite a bit of data without using more words. Additionally, the graphic does not contain any redundant or repetitive data.

Gestalt principles

Since the lines display the same data for different transmissions and as each line displays a specific transmission in its own colour, this graph corresponds to the concepts of proximity and similarity (of colour). This graph adheres to the continuity principle since the price increase is consistent with the increase in the year, but it does not adhere to the closure and connection principles because the objects in the graph cannot be seen as complete objects and they also cannot be connected.

The graph orientation also follows IBCS rules.

Section 3 (Descriptive Statistics)

The words "mean," "median," and "mode" refer to the dataset's average value for each given variable, its median value, and its most prevalent value, respectively.

After selecting the sample dataset, we are carrying out a descriptive analysis as part of which we will identify the majors of central tendency and majors of spread.

1) Majors of spread is composed of variance and standard deviation

2) The main components of mean, median, and mode central tendency.

There are different sections in the dataset which consists of some variable being numerical data while others being non-numerical data. The summary of descriptive statistics in table is shown below.

The variance of a dataset is the difference between the average and the individual values (for instance the average of the car price of that dataset and car price of a single car)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
price	99	11995	44995	20755.92	6419.048	41204181.463	1.247	.243	1.334	.481
year	99	0	5	3.43	1.117	1.248	-.840	.243	.433	.481
mileage	99	50	120000	42397.38	23022.200	530021693.99	.608	.243	.562	.481
engine	99	1.3	2.1	1.552	.2753	.076	.874	.243	-.653	.481
power	99	108	381	157.41	61.762	3814.572	1.959	.243	2.878	.481
Valid N (listwise)	99									

The variance of a dataset is the difference between the average and the individual values (for instance the average of the car price of that dataset and car price of a single car). Also, now we will be comparing the non-numeric and numeric variables from the dataset which are fuel type and car price.

fuel type	count	mean	std	min	25%	50%	75%	max
diesel	36	17391.81	3549.107	11995	14823.75	17099	18757.5	28318
petrol	62	22615.84	6944.377	14649	17849	20299	28373	44995

We have taken into account the numeric variable of car price and the non-numeric variable of fuel type. Now that we are comparing both of these factors, we can infer that petrol cars are preferred more than diesel automobiles on the basis of the values we acquired. You can see this in the above table.

Section 4 (Confidence Interval)

Now, assuming that the average price of a used car is normally distributed, which means that the data will be spread about the centre value. Also, now using confidence interval of sample mean, we will determine population mean with 95% confidence, which means that out of 100 times, as the confidence interval is mean of our estimate

and minus the difference in that approximate number, we are confident that the estimate will fall between the lower and upper values specified by the confidence interval 95 times.

Sample mean = 20755.92

Population interval 20628.5359<p<20883.0041 (where p = population mean)

so, the average population of car price lies between 20628.5359 and 20883.0041

Section 5 (Null Hypothesis Testing)

One-Sample Test

Test Value = 21414.66

	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	-1.021	98	.155	.310	-658.741	-1939.00	621.52

We now compare the average vehicle price based on our sample data and the average car price of that particular car model with the UK auto market.

Ho(mean of sample - mean of population) = 0

Ha != 0

The average price of second-hand car is 21414.66£ while the average price of this second hand car in UK market is 24312 £.

Source: carsite.co.uk

We may compare the average cost of used cars and arrive at the conclusion that they are not that close to the average cost of the UK market. To determine if this is pass or fail, we will apply the hypothesis testing. To verify, we will perform a one sample t-test.

The value for P we get after conducting the test is 0.155

Therefore we can now conclude that there is a significant difference between the average car price from our sample and average car price of Mercedes a-class in the UK market.

Section 6 (Pearson correlation)

The most typical technique to determine a linear correlation is the Pearson correlation coefficient(r). This approach is the statistical test which measures the statistical relationship, between two numeric variables. The strength and direction of the link between two variables is expressed as number between -1 and 1 and it is known as (r), we can write it as -1<r<1.

There are two types of r values, positive and negative correlation, in negative correlation if the age of car decreases with the increase in age of car while in positive correlation , while in positive correlation if one variable increases will result into other variable change , for instance, in sample dataset, with increase in price the engine size increases.

Considering the below matrix. We can observe that as the older the car gets, the price of car decreases, it shows us that there is negative correlation between these two variables this way we can see that the strength of correlation between this two variables is r = -0.707.

Correlations

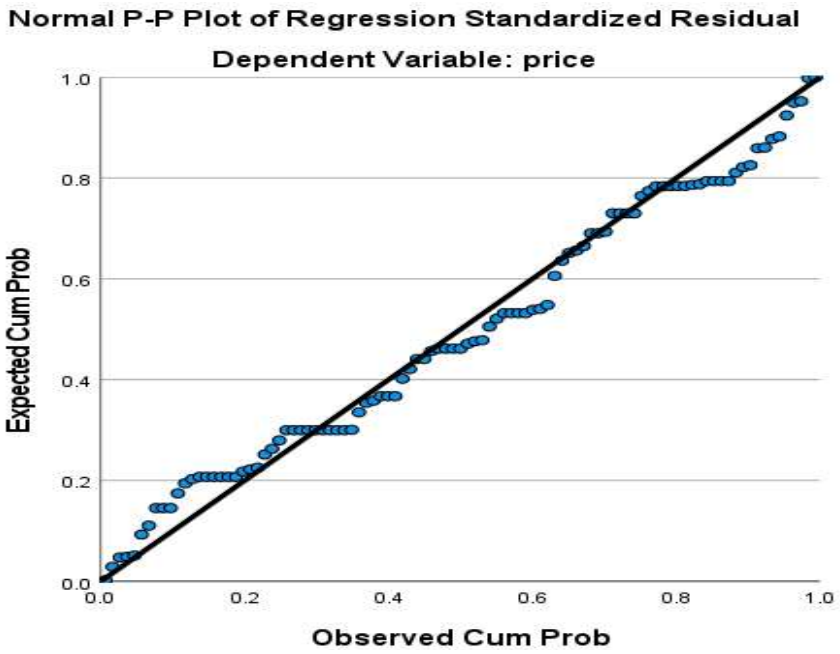
		price	year	mileage	power	trasmission_type	fuel_type
Pearson Correlation	price	1.000	-.707	-.565	.767	-.464	.402
	year	-.707	1.000	.519	-.344	.271	-.379
	mileage	-.565	.519	1.000	-.232	.167	-.348
	power	.767	-.344	-.232	1.000	-.272	.471
	trasmission_type	-.464	.271	.167	-.272	1.000	.025
	fuel_type	.402	-.379	-.348	.471	.025	1.000

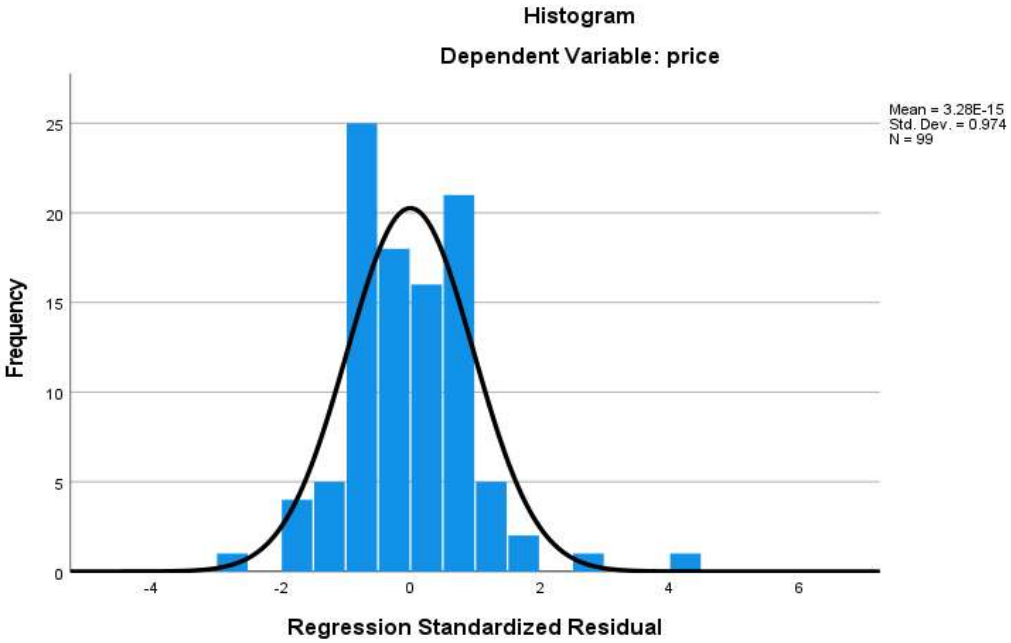
we can also conclude that the impact of fuel type category will be less as compared to other variables while purchasing a second-hand car, this will vary in actual conditions as we can perform regression analysis.

Section 7 (Regression Analysis)

The model which fits the data properly is known as parsimonious model which is also on the other hand more simple and needs only few parameters which is better than a detailed and a complex model.

To see how well it pairs a set of observations, a statistical model's goodness of fit is used to measure. We can also derive that the graph for this is the parsimonious model as the it fits well with the data and we can also see that it plotted as almost y = x line.





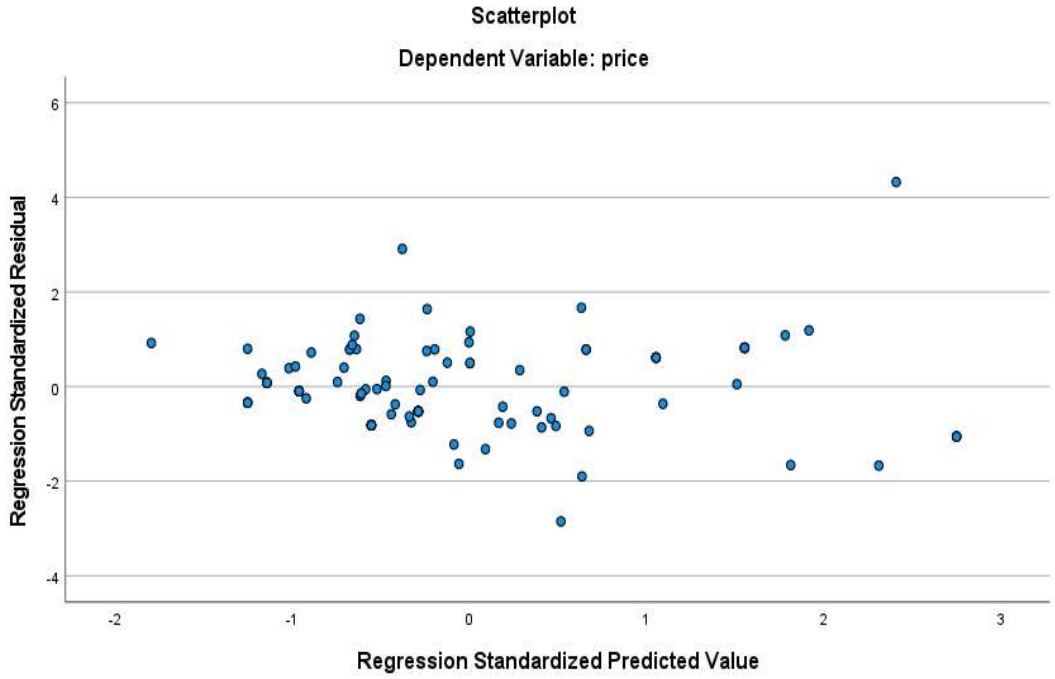
We can also conclude from the above histogram that the number of car with prices less than mean are greater than the number of cars which are less than the mean price.

Section 8 (Residual Analysis)

Now we will look at the scatter graph, from this graph we can assert that linearity is met if there are points or values that are randomly dispersed or widely spread.

The absence of trends in the residuals of the graph suggests that it adheres to the homoscedasticity and independence of error principles. Also, the ranges in which the majority of residuals are randomly distributed is -2 and 2, and 0 and -2.

Now, we can say that the model is adequate as all the assumptions are satisfied.



Section 10 (Regression result)

We now know the value of Adjusted R square is .878 and we also know that the model is adequate and parsimonious.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.940[a]	.884	.878	2239.865

a. Predictors: (Constant), fuel_type, trasmission_type, mileage, power, year

b. Dependent Variable: price

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	29878.677	1995.989		14.969	<.001
	year	-2142.621	254.318	-.373	-8.425	<.001
	mileage	-.067	.012	-.242	-5.744	<.001
	power	60.553	4.420	.583	13.700	<.001
	trasmission_type	-2315.565	553.791	-.161	-4.181	<.001
	fuel_type	-1199.374	557.109	-.094	-2.153	.034

a. Dependent Variable: price

In order to achieve parsimonious model, we will eliminate those variables who value in the coefficient table is greater than 0.05.

With increase in price

We also know that,

This model is 87% accurate.

Conclusion

From the findings it was found that the mileage and the age had significant impact on the price of the car while the transmission type and fuel type had comparatively little impact on the price of 2nd hand car.

Therefore, the price calculator for our model will be:

Price = constant + year*(-2142.621)+mileage(-0.67)+power(60.553)+transmission_type(-2315.565)