Objectives of the report:

We conduct a study on the second-hand automobile market in the UK for a particular model and make, on behalf of a market research business working for Car4all. We assist the marketing research agency in identifying the most significant factors in determining a used car's pricing. We further develop a statistical model to forecast a car's market value solely based on its mileage, engine size, and other optional extra factors that affect the car's selling price.

Source and Description:

The research examines the costs of used Volkswagen Passat vehicles in Stanmore. "www.autotrader.co.uk", a respected and specialized website, was used to collect the necessary data. The Volkswagen Passat's selling price, miles-driven, body design, Transmission type, and fuel type are all the elements investigated further. 300+ different data values were scraped from the website and a sample of 118 rows was chosen for further study.

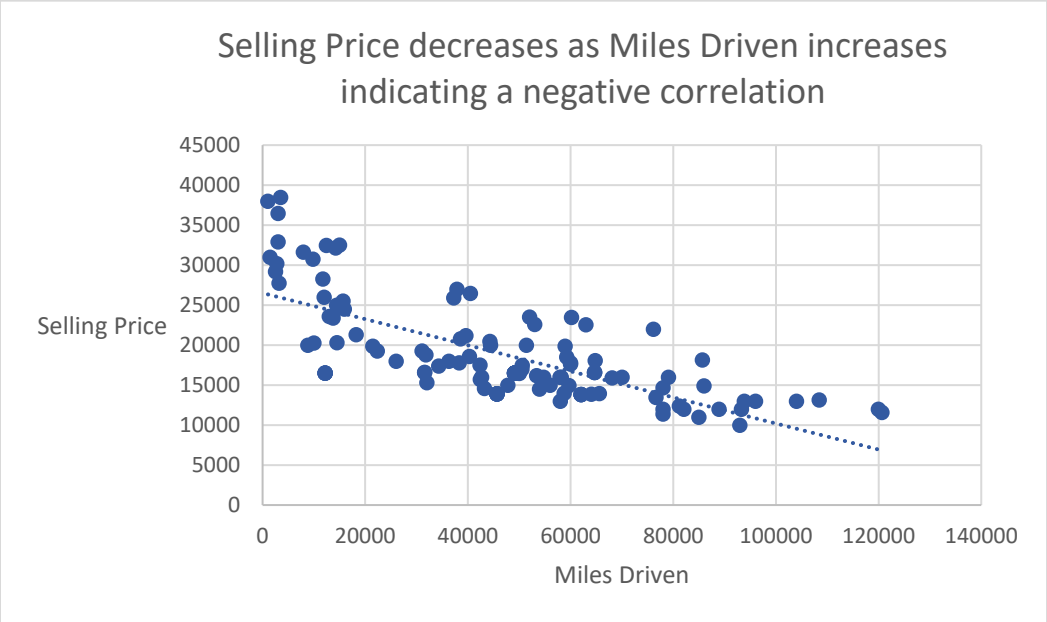
Challenges: The data was acquired via the internet, and then it was wrangled to deal with unstructured data. Our observations suggested that values were missing at random, which led us to conclude that the data sample was still representative of the population. For further analysis, missing values were eliminated from the dataset.

Sampling Method and Data representativeness: A simple random sampling method is used to extract the samples from the population. A simple random sample is a subset of a statistical population with an equal chance of being chosen. A simple random sample reflects a group in an unbiased way. Because every member of the population has an equal probability of being chosen, it is a fair technique to pick a sample from the population dataset. As a result, we can say that our sample is representative of the population dataset.

Consideration of 5-year range of registration for the cars:

When we look at cars manufactured more than five years ago, the selling price is significantly lower than a car that was just registered or manufactured (also affected by other factors). Including these kinds of data, outliers would arise in the analysis resulting in skewed or biased results. By deleting or not considering these sorts of numbers and developing the most parsimonious model possible, we hope to reduce the impact of outliers.

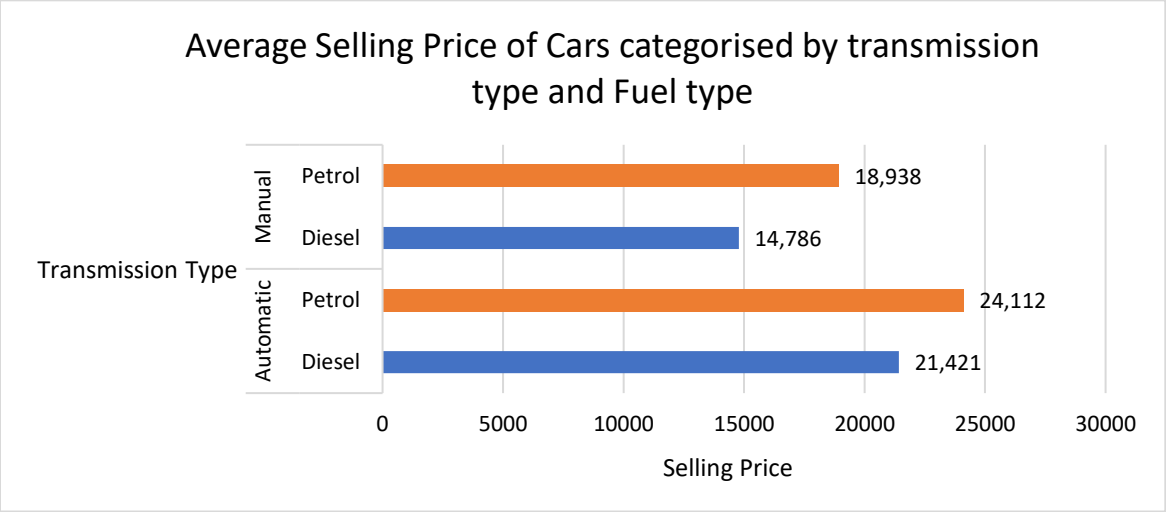
Visualization:



Scatter Plot: This graph is used to examine if two variables have a relationship or association.

The graph clearly illustrates a negative relationship between the selling price and the number of miles driven. The value of a car decreases as the number of miles driven rises. In other words, automobiles with fewer miles on the clock demand a higher premium than cars with more miles on the clock.

The line gives the direction of correlation. The graph displays the relationship between the variables clearly and concisely. The graph has a clear context and just a few design variations. The background is visually appealing and assists with ink conservation.

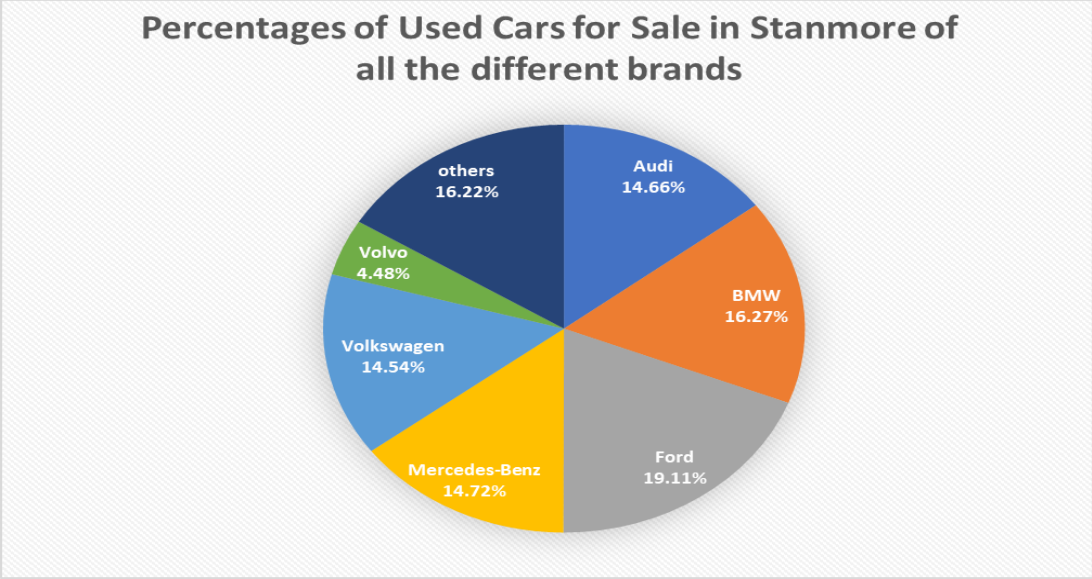


A bar chart (sometimes known as a bar diagram) is a visual tool that uses bars to compare data between categories. A bar graph can be oriented horizontally or vertically.

It's employed here as it enables comparing things in different groups easier.

The graph above shows the average selling price of automobiles by gearbox type and fuel type, which are categorical variables. Automatic transmission petrol car owners demand a higher resale value for their vehicles. It's logical to assume that used automatic autos are more expensive than used manual cars. Manual diesel cars are the cheapest, with an average price of 14786 pounds.

The graph shows that it is proportional as the y-axis is 0 and the vertical scale is linear. The title and data labels are all stated clearly. It's a simple data-driven two-dimensional graph with few design variations. The information is conveyed clearly and concisely. With color similarity, it is easy to distinguish between various selling prices. Appropriate spacing is maintained to differentiate between categories of cars.



Pie Chart: The percentages represented by each category are displayed next to the proper slice of pie in this pie chart to represent percentage or proportional statistics.

The pie chart depicts the percentage of used cars offered for sale in Stanmore of various makes in the population. When compared to all other manufacturers, Ford has the most used vehicles for sale, accounting for 19.11 percent of the total used car population on sale. Volvo, on the other hand, has the fewest used automobiles for sale.

The numerical amounts depicted are directly proportionate to the proportions (pies). The graph uses clear, precise, and extensive labeling.

Descriptive Statistics:

Row Labels	Count of Selling Price	Average of Selling Price	Standard Deviation of Selling Price	Max of Selling Price	Min of Selling Price
Automatic					
Diesel	33	21420.84424	7335.865393	37980	12000
Petrol	25	24112.2352	5504.125695	38490.44	16585.44
Manual					
Diesel	53	14785.50075	2210.488608	20490	10000
Petrol	7	18937.63	1859.398198	21195.44	16600.44

The table above shows descriptive information for the Volkswagen Passat car's selling price when broken down by transmission and fuel type categories. In comparison to other types of diesel cars, manual transmission diesel cars are in great demand. We may deduce from the preceding data that, regardless of the fuel burned, the average selling price of automatic cars is greater than that of manual transmission cars. For an automatic transmission petrol car, a higher maximum selling price was demanded than for others. The selling price range in which used cars are sold may be calculated using the highest and minimum selling prices.

Confidence Intervals:

Transmission Type	Fuel Type	Average Selling Price	Standard Deviation of Selling Price	Count	Confidence Intervals (95%)		Margin Of Error
					Lower Bond	Upper Bond	
Automatic	Petrol	24112.2352	5504.125695	25	21954.7	26270	2157.58
	Diesel	21420.84424	7335.865393	33	18918	23923.7	2502.9
Manual	Petrol	18937.63	1859.398198	7	17560.2	20315	1377.44
	Diesel	14785.50075	2210.488608	53	14190.389	15380.611	595.111

95 percent confidence intervals of the average second-hand automobile selling price are presented above after categorizing them by gearbox and fuel type, assuming that second-hand car prices are normally distributed.

"A confidence interval is a set of values derived from sample data that are most likely to contain the value of an unknown population parameter."

For example, if we look at the Volkswagen Passat automatic diesel automobile, we can see from the table above that CI= [18918,23923.7] at a 95% confidence level. This suggests that if you compute the average selling price over the whole population, the average will fall within this confidence interval. And we're almost positive about that.

Hypothesis Testing:

	Average Selling Price As per Sample Data	Average Selling in the UK	Mean Difference	Count n
Automatic				
Diesel	21420.84424	21779	258.15	33
Petrol	24112.2352	25184	1071.76	25
Manual				
Diesel	14785.50075	15156	370.49	53
Petrol	18937.63	19660	722.37	7

A hypothesis test is used to determine if the average price of a used Volkswagen Passat car from our data sample is comparable to the average price in the UK.

The sample mean under test=14785.50075(Manual Diesel)

The average selling price of the car in the UK is referred from: www.autotrader.co.uk

Test performed to compare with the average price in the UK= One Sample two-tailed T-test

One-Sample T-test of means compares the sample mean to a hypothesized value of the population mean and looks for deviations.

Null Hypothesis(H0): 'No' significant difference between the sample mean and the hypothesized mean of the population (sample mean=population mean).

Alternate Hypothesis(H1): Significant difference between the sample mean and the hypothesized mean of the population (sample mean != population mean).

After conducting the test at a 95% confidence level, the p-value obtained is:>0.05

As a result, we fail to reject the Null Hypothesis since the p-value is greater than the level of significance (0.05). In another way, the difference between the average selling price derived from sample data and the population parameter is not statistically significant.

Correlation Matrix:

Correlations							
		Price	Miles Driven	Transmission Type	Body Design	Fuel Type	NO Of Previous owners
Pearson Correlation	Price	1.000	-0.728	-0.591	0.191	-0.406	0.057
	Miles Driven	-0.728	1.000	0.346	0.138	0.297	0.030
	Transmission Type	-0.591	0.346	1.000	-0.237	0.354	-0.152
	Body Design	0.191	0.138	-0.237	1.000	0.017	0.194
	Fuel Type	-0.406	0.297	0.354	0.017	1.000	-0.086
	No of Previous owners	0.057	0.030	-0.152	0.194	-0.086	1.000

Correlation: A statistical measure of the strength and direction of a link between two or more variables.

The Pearson correlation coefficient is used to determine the strength of the association between the variables.

The table shows the miles driven, transmission type, and fuel type all have a negative relationship with price, but body design and the number of previous owners have a positive relationship.

Statistical conclusions will be reliable since the independent variables are not multicollinear.

Regression: is a statistical method for determining the degree and nature of a relationship between one dependent variable (typically indicated by Y) and a set of independent variables (known as independent variables).

The dependent variable (Y) in our model is: Selling Price

The independent variables (Xi) are: Miles Driven, Transmission Type, Body Design, Fuel Type, no of previous Owners

SPSS software is used to carry out the analysis.

Initially, all the above-mentioned independent variables were used to build the regression model. But the significance value for the No of Previous Owners was more than 0.05(which indicates that it has no statistically significant impact in determining the selling price). Hence it was dropped from the model and the model was rebuilt with the remaining independent variables to get the parsimonious model. The model now contains only those independent variables that have a significant impact on the dependent variable as their p-value is less than 0.05(almost nearing '0').

Model Summary:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.843[a]	.711	.700	3398.680	.711	69.357	4	113	<.001

a. Predictors: (Constant), Body Design, Fuel Type, Miles Driven, Transmission Type

b. Dependent Variable: Price

The model's adjusted R square is 0.700.

According to the Adjusted R Square, the independent variable (Selling Price) accounts for 70% of the variation in the dependent variable (Selling Price) (Miles Driven, Transmission Type, Body Design, Fuel Type).

This indicates that the model is effective and suitable for analysis.

Residual Analysis:

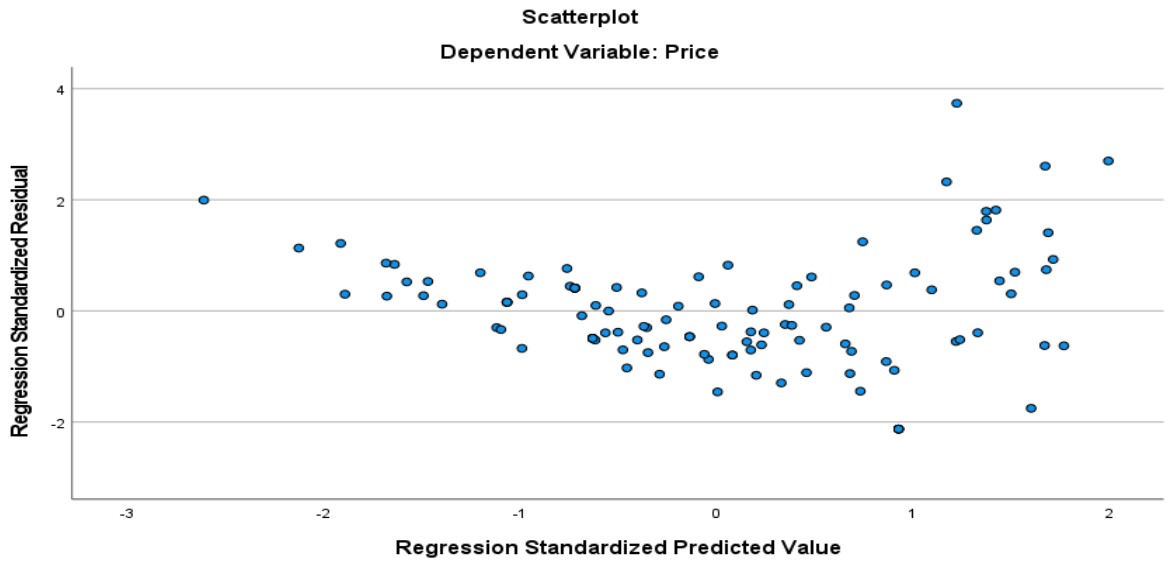
Used to check if a simple or multiple regression has succeeded in explaining as much variance in a dependent variable as possible while remaining faithful to the underlying assumption. All residuals should ideally be small and unstructured, suggesting that the regression analysis successfully explained the bulk of the variance in the dependant variable.

Assumption 1: The average of the residuals must equal zero, implying that the residuals must be dispersed uniformly around zero.

We can see that assumption 1 is met by looking at the scatter graph because the residuals are uniformly distributed.

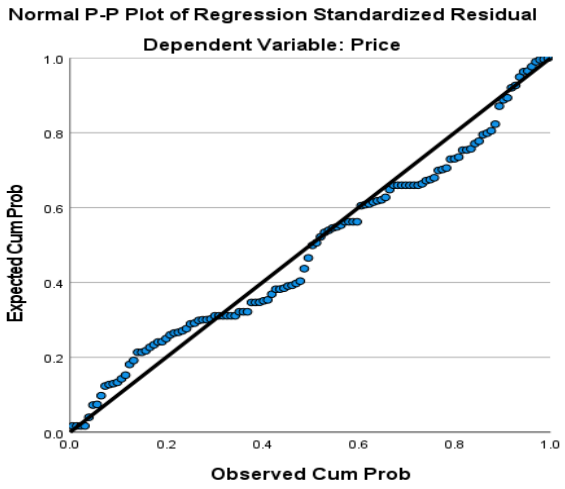
Assumption 2: There is no discernible pattern between the residuals, or the residuals are unrelated.

The graph has no discernible trends, and the residuals are distributed around '0.' As a result, assumption 2 is correct.



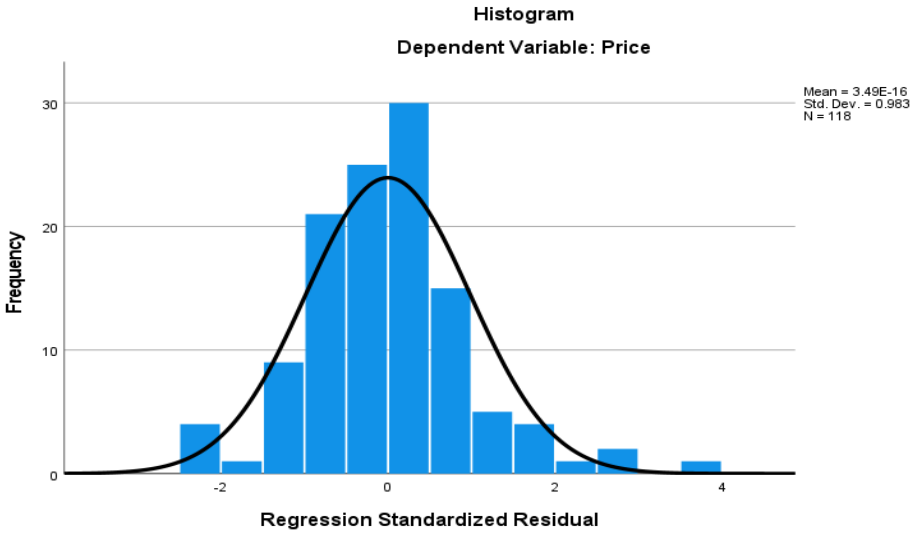
Assumption 3: There is no significant difference between the standardized residuals and the residuals.

The validation is done using a Normal P-P plot. The least variation between standardized residuals and the best fit line may be seen in the graph below.



Assumption 4: The residuals follow a normal distribution.

We created a histogram of the residuals to verify this. The residuals are regularly distributed, according to the assumption, as seen by the graph. The histogram is virtually identical to the normal distribution graph's bell curve. As a result, assumption 4
is met.



Assumption 5: Multicollinearity does not exist.

When independent variables are highly related to one another, multicollinearity occurs. When we look at the correlation matrix that we created before, we can see that the independent variables do not have multicollinearities. As a result, assumption 5 is met.

The histogram looks approximately normal, and the scatterplot of residuals are randomly scattered. Hence, we can conclude that, the model is appropriate.

Derived statistical model:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	27173.964	769.117		35.331	<.001
	Miles Driven	-.139	.013	-.622	-11.016	<.001
	Transmission Type	-3473.442	731.382	-.281	-4.749	<.001
	Fuel Type	-1752.934	769.328	-.126	-2.279	.025
	Body Design	2629.023	665.588	.213	3.950	<.001

a. Dependent Variable: Price

Our model is thus described by the line:
Selling Price=27173.964 -0.139∗MilesDriven− 3473.442∗TransmissionType+ 2629.023∗BodyDesign− 1752.934∗FuelType

Based on the indications on the coefficients, we may conclude that Body Design has a positive impact on the Selling Price, whereas Miles Driven, Transmission Type, and Fuel Type have a negative impact.
To determine the selling price, examine the following values:
Miles Driven:31083
Transmission Type: Automatic (0)
Body Design: Saloon (0)
Fuel Type: Diesel (1)
The predicted Selling price is: 15505.553 pounds.