DESCRIPTIVE ANALYTICS

**Table of Contents**

Executive Summary…...........................................................................................................2

Objectives of the report..........................................................................................................2

Description.............................................................................................................................2

Source and Issues...................................................................................................................2

Sampling…………………….……………………………………………………….…......2

Visualization..........................................................................................................................2

Bar Chart…………....................................................................................................3

Scatter Chart……......................................................................................................4

Pie Chart....................................................................................................................5

Descriptive statistics..............................................................................................................6

Confidence intervals……………………………………………… ……………………….7

Hypothesis testing………………………………………………………………………….7

Correlation Analysis…………………………………………………………………….….8

Regression Analysis……………………………………………………………………..…9

Residual Analysis.................................................................................................................11

Application of the Model………………………./………………………………………...14

References............................................................................................................................15

**INTRODUCTION:**

This report was created in order to conduct market research on a used car of a specific make and model on behalf of car4all focusing on the important variables affecting the price of second-hand car.

**Source:**

Data collection for a specific make and model is done to conduct research on the used car market. The website that was used to collect the data is <https://www.autotrader.co.uk/cars/used>. For this research, we have chosen the brand MERCEDES BENZ and the model B class. The last five-year data (from 2017 to 2022) was collected because the price before that would be in a less range compared to the latest data. In total, we have 341 data entries with the corresponding variables.

**Sampling of data:**

Using the sampling method, we randomly selected 116 data points from the available 341 data points. Later, the data was cleaned up by dealing with missing values and removing unwanted values.

**Goal:**

The goal is to determine which variables are important to consider when developing a statistical model of a car to estimate market value and how each variable affects the price of these cars. In this report, we have chosen a specific sample from the data that we have scraped from the previously mentioned website. We are ***visualizing*** the data by using the appropriate variables while ensuring the principles and adequateness by adhering to the guidelines. Then we have analyzed the data by performing ***descriptive analysis, correlation analysis, confidence interval analysis and regression analysis.***

**Description:**

The variables that we have chosen to analyze the secondhand car market is ‘Price’ which is the dependent variable whereas ‘Year’, ‘Mileage’, ‘Engine capacity’, ‘Engine type’ and ‘Fuel type’, ‘no of doors’ and ‘no of seats’ are the independent variables. We are determining the interrelationship of these independent variables with the dependent variable.

**VISUALIZATION:**

We visualized the data using different variables to show the interrelationship between them and represented it in three different types of charts.

1. **Bar chart:**

Chart, bar chart

Description automatically generated

Figure 1: Bar chart of average of price by model and year

*Evaluation:*

Bar chart represents the categorical variables with respect to the axis either vertically or horizontally. In this Figure 1 we have plotted.

This bar chart follows similarity and enclosure of gestalt principles. The similarity principle is maintained in this graph as each year of registration has the same color. We can also say that this graph follows enclosure as it is enclosed in the particular blocks with one color for one value of the year.

This graph demonstrates proportionality from Tufte’s principle as the scale starts from zero. We can also say this bar chart is data driven as each block size in the graph represents the size of the data present in the data. This also represents the appropriate dimensions as the information carrying dimensions did not exceed the number of dimensions in the data.

1. **Scatterplot:**

Price in pound

Figure 2: Correlation between price and mileage

Continuity, proximity,connection, similarity

*Evaluation:*

Figure 2 represents the correlation between price and miles in this scatterplot. The x-axis represents the price in pounds and the y-axis represents the mileage in thousands. This line shows that the price is negatively correlated to mileage.

This scatterplot represents the proximity of gestalt principles as the plot shows the data closely plotted with each other. It also represents the similarity and continuity principle as the exponential line of this scatterplot is plotted. Similarity is followed as the data plots are represented as they are more closely related near the range 15,000 to 30,000.

This scatterplot shows the clarity of Tufte’s principle as the exponential line in this graph shows a clear picture of the price and the mileage.

1. **Pie chart:**

Chart, pie chart

Description automatically generated

Figure 3: Pie chart for representing the average of price by engine capacity

*Evaluation:*

Here Figure 3 is a pie chart which depicts the average of price along with their percentage with the engine capacity.

This graph follows proportionality, contextualized and data driven aspects of the Tufte’s principle as it represented in two dimensions. Contextualized is represented in this graph as it displays the appropriate data as the original data.

**DESCRIPTIVE STATISTICS TABLE:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | | | | | | | |
|  | N | Range | Minimum | Maximum | Sum | Mean |  | Std. Deviation | Variance | Skewness |  | Kurtosis |  |
|  | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| price | 116 | 23516.00 | 12479.00 | 35995.00 | 2419266.00 | 20855.7414 | 510.83510 | 5501.86241 | 30270489.967 | 0.823 | 0.225 | -0.110 | 0.446 |
| mileage | 116 | 84858.00 | 1014.00 | 85872.00 | 2752409.00 | 23727.6638 | 1752.29206 | 18872.76307 | 356181185.964 | 1.436 | 0.225 | 1.807 | 0.446 |
| Engine capacity | 116 | 0.80 | 1.30 | 2.10 | 199.40 | 1.7190 | 0.02829 | 0.30472 | 0.093 | 0.085 | 0.225 | -1.588 | 0.446 |
| age | 116 | 4.00 | 1.00 | 5.00 | 426.00 | 3.6724 | 0.12045 | 1.29733 | 1.683 | -0.558 | 0.225 | -0.882 | 0.446 |

Table 1: Descriptive statistics table for the continuous variable

We are summarising the continuous variable by using the descriptive statistics table. The maximum and minimum values of all variables, as well as the mean, standard deviation, and sample size, are shown in this table. The highest price is 35995, and the lowest price is 12479. Descriptive statistics is obtained from SPSS software.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Average of price | Std Dev of price | Count of price | Max. of price | Min. of price |
| Automatic | 21657.2 | 5621.02 | 96 | 35995 | 13000 |
| Diesel | 19961.9 | 4881.0329 | 64 | 31719 | 13000 |
| Petrol | 24351.9 | 4968.241 | 30 | 35099 | 16750 |
| Petrol Plug-in Hybrid | 35487.5 | 717.71338 | 2 | 35995 | 34980 |
| Manual | 17008.6 | 2532.0178 | 20 | 21500 | 12479 |
| Diesel | 14476.3 | 2087.1733 | 7 | 17399 | 12479 |
| Petrol | 18372.1 | 1489.3287 | 13 | 21500 | 16000 |

Table 2: Descriptive statistics table by grouping ‘fuel type’ and ‘engine type’

The descriptive statistics for the categorical variable are summarised in this table. In this case, two categorical variables, 'engine type' and 'fuel type,' are chosen, with price as the dependent variable.

**CONFIDENCE INTERVALS (95% AND 99%):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **CONFIDENCE INTERVALS (in pounds)** | | |  |
|  | 95% |  | 99% |  |
|  | LOWER BAND | UPPER BAND | LOWER BAND | UPPER BAND |
| Automatic |  |  |  |  |
| Diesel | 18,766.09 | 21,157.75 | 18,390.33 | 21,533.51 |
| Petrol | 22,574.07 | 26,129.73 | 22,015.44 | 26,688.36 |
| Petrol Plug in Hybrid | 34,492.82 | 36,482.18 | 34,180.27 | 36,794.73 |
| Manual |  |  |  |  |
| Diesel | 12,930.11 | 16,022.46 | 12,444.27 | 16,508.30 |
| Petrol | 17,562.48 | 19,181.67 | 17,308.09 | 19,436.06 |

Table 3: Confidence intervals table

The above table shows the values of the 95% and 99% confidence intervals for the average price range for the engine types with respect to the fuel type.

From the above table we can be 95% confident the value of Automatic Diesel vehicles lies in the range 18,766.09 to 21,157.75. Similarly for the other types we are confident that the price value lies in that particular range.

From the above table we can be 99% confident the value of Automatic Diesel vehicles lies in the range 18,390.33 to 21,533.51. Similarly for the other types we are confident that the price value lies in that particular range.

**ANALYSIS OF THE MEAN DATA - HYPOTHESIS TESTING:**

We are performing hypothesis test in order to analyse the mean of the sample we have with the whole population.

**Null hypothesis(H0):** The mean of the sample data is equal to the mean of the population taken from the website.

**Alternative Hypothesis(Ha):** The mean of the sample is not equal to the mean of the population taken from the website.

|  |  |  |  |
| --- | --- | --- | --- |
| Mean price of the sample | Average price in UK | Standard deviation | Sample T test value(p value) |
| 20855.7414 | 19956 | 5501.86241 | 0.0808 |

Table 4: T test value

As the p value is greater than 0.05 so we accept null hypothesis. After conducting this t test we can now say that the mean of the sample data is equal to the mean of the population taken from the website.

**CORRELATION ANALYSIS:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | CORRELATION | |  | |  |  |  |
|  | PRICE | MILEAGE | PETROL | PETROL\_PLUGIN | EC1.3 | EC1.5 | EC1.6 | EC2 | AGE |
| PRICE | 1 | -0.646 | -0.237 | -0.354 | -0.632 | 0.39 | 0.191 | -0.434 | -0.903 |
| MILEAGE | -0.646 | 1 | 0.384 | 0.124 | 0.338 | -0.456 | 0.18 | 0.234 | 0.54 |
| PETROL | -0.237 | 0.384 | 1 | -0.102 | 0.5 | -0.382 | 0.683 | -0.34 | 0.275 |
| PETROL\_PLUGIN | -0.354 | 0.124 | -0.102 | 1 | 0.29 | -0.066 | -0.069 | -0.059 | 0.223 |
| EC1.3 | -0.632 | 0.338 | 0.5 | 0.29 | 1 | -0.227 | -0.239 | -0.202 | 0.556 |
| EC1.5 | 0.39 | -0.456 | -0.382 | -0.066 | -0.227 | 1 | -0.261 | -0.22 | -0.294 |
| EC1.6 | 0.191 | 0.18 | 0.683 | -0.069 | -0.239 | -0.261 | 1 | -0.232 | -0.117 |
| EC2 | -0.434 | 0.234 | -0.34 | -0.059 | -0.202 | -0.22 | -0.232 | 1 | 0.465 |
| AGE | -0.903 | 0.54 | 0.275 | 0.223 | 0.556 | -0.294 | -0.117 | 0.465 | 1 |

Table 5: Correlation table

In order to perform correlation between the variables we need to convert the categorical variables into numerical values. Here we are converting the ‘year’ into age of the car. The other categorical variables like engine type, engine capacity, fuel type are converted into numerical variables by using dummy variables. The correlation is then done between all the variables irrespective of they are dependent or independent. In this table the variable age is negatively correlated with price. Also mileage is negatively correlated with the dependent variable ‘price’. From this table we can infer that the engine capacity 1.5 is highly correlated with price which tells us that price actually depends on this variable type. Another interesting correlation is the fuel type petrol is positively correlated with petrol and negatively correlated with the petrol plug in hybrid type.

Price is negatively correlated with mileage and engine capacity 1.3litres, indicating that these variables will have a negative impact on price.

SPSS software was used to generate this correlation matrix.

**REGRESSION:**

In order to perform regression we used SPSS software. We did linear regression analysis in SPSS.

We perform linear regression analysis on the dataset to determine the dependence of the variables. In our case, the dependent variable is price, which can vary depending on the other independent variables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Coefficients** | |  |  |
|  |  | Unstandardized Coefficients |  | Standardized Coefficients | t | Sig. |
|  |  | B | Std. Error | Beta |  |  |
|  | (Constant) | 39410.472 | 1649.869 |  | 23.887 | 0 |
|  | MILEAGE | -0.063 | 0.011 | -0.217 | -6.008 | 0 |
|  | PETROL\_PLUGIN | -6052.697 | 1219.716 | -0.144 | -4.962 | 0 |
|  | EC1.3 | -3101.565 | 752.435 | -0.214 | -4.122 | 0 |
|  | EC1.5 | 911.391 | 475.09 | 0.066 | 1.918 | 0.058 |
|  | EC1.6 | 1187.032 | 490.047 | 0.089 | 2.422 | 0.017 |
|  | EC2 | -2226.274 | 720.581 | -0.15 | -3.09 | 0.003 |
|  | AGE | -2270.393 | 223.622 | -0.535 | -10.153 | 0 |

Table 6: Regression table (insignificant variable: Engine capacity 1.5)

The significance value of the engine capacity (1.5L) is 0.058 in this linear regression analysis, which is greater than 0.05. As a result, we conclude that this variable is insignificant and must be removed because it makes the model less parsimonious.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Coefficients** |  |  |  |
|  | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
|  | (Constant) | 41097.096 | 1413.15 |  | 29.082 | 0 |
| MILEAGE | -0.068 | 0.01 | -0.234 | -6.611 | 0 |
| PETROL\_PLUGIN | -6095.485 | 1234.413 | -0.145 | -4.938 | 0 |
| EC1.3 | -3610.338 | 712.749 | -0.249 | -5.065 | 0 |
| EC1.6 | 812.955 | 455.073 | 0.061 | 1.786 | 0.077 |
| EC2 | -2738.749 | 677.409 | -0.185 | -4.043 | 0 |
| AGE | -2175.634 | 220.764 | -0.513 | -9.855 | 0 |

Table 7: Regression table 2 (insignificant variable: Engine capacity 1.6)

The variable engine capacity (1.6L) makes the model insignificant in this analysis, as the significance value is 0.077, which is greater than 0.05. To keep the model as simple as possible, we remove this variable.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Coefficients** |  |  |  |
|  | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
|  | (Constant) | 41097.1 | 1413.15 |  | 29.082 | 0 |
| MILEAGE | -0.068 | 0.01 | -0.234 | -6.611 | 0 |
| PETROL | 812.955 | 455.073 | 0.072 | 1.786 | 0.077 |
| PETROL\_PLUGIN | -5282.53 | 1332.746 | -0.126 | -3.964 | 0 |
| EC1.3 | -4423.293 | 639.007 | -0.305 | -6.922 | 0 |
| EC2 | -2738.749 | 677.409 | -0.185 | -4.043 | 0 |
| AGE | -2175.634 | 220.764 | -0.513 | -9.855 | 0 |

Table 8: Regression table 3(insignificant variable: Petrol)

The significance value of petrol is 0.077. We remove this variable to maintain the significance of the model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Coefficients** |  |  |  |
|  | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
|  | (Constant) | 42294.33 | 1256.456 |  | 33.662 | 0 |
| MILEAGE | -0.062 | 0.01 | -0.212 | -6.329 | 0 |
| PETROL\_PLUGIN | -6184.484 | 1245.632 | -0.147 | -4.965 | 0 |
| EC1.3 | -4212.47 | 634.239 | -0.29 | -6.642 | 0 |
| EC2 | -3303.751 | 604.981 | -0.223 | -5.461 | 0 |
| AGE | -2081.811 | 216.551 | -0.491 | -9.614 | 0 |

Table 9: Regression table 4

We have now eliminated all of the insignificant variables. This model is now a parsimonious model.

We can conclude that the variables MILEAGE, PETROL PLUGIN, EC1.3, EC2, and AGE have an effect on the PRICE variable.

We can obtain the equation of PRICE as:

**PRICE = 42294.329 - (0.062 \* MILEAGE) - (6184.484 \* PETROL\_PLUGIN) - (4212.47 \* ENGINE\_CAPACITY\_1.3) - (3303.751 \* ENGINE\_CAPACITY\_2) - (2081.811 \* AGE)**

**RESIDUAL ANALYSIS:**

Residual analysis is performed in SPSS software to find how appropriate the model is by defining the residual points and by plotting the required graphs. We perform residual analysis to find the difference between the observed and the predicted values. We take four assumptions in consideration.

***Assumption 1:***

In the first assumption we assume that the residual values are normally distributed over zero. Here in this scatterplot the residual is spread across zero hence the assumption proves to be true.

Chart, scatter chart

Description automatically generated

Figure 4: Residual scatterplot

***Assumption 2:***

In the second assumption we assume that the residuals need to be independent. Here we can prove the assumption as the residuals do not follow any pattern above or below zero. This assumption does not hold true when either the residuals are grouped or they appear in a pattern.

***Assumption 3:***

Here we assume that the standard deviation of the dependent variable is same for all the residuals. We can prove that as it deviates from the linear line in this normal P-P plot below.

Chart, line chart, scatter chart

Description automatically generated

Figure 5: Normal p-p plot of the standard deviation of dependent

***Assumption 4:***

Assumption 4 states that the residuals are normally distributed. We can prove this assumption by looking at the below histogram which states that the residuals are normally distributed.

Chart, histogram

Description automatically generated

Figure 6: Normal distribution of the dependent variable

***Assumption 5:***

In this last assumption we assume that multicollinearity doesn’t exist. We can prove this assumption by looking at the below table where we have to prove that no independent variables have multicollinearity exists between them as the values are between 0.7 to -0.7. Here in the below table no values of correlations are either greater than 0.7 or less than -0.7. Hence this assumption is also proved`

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | COEFFICIENT CORRELATIONS | | |  |
|  | | | AGE | PETROL\_PLUGIN | EC2 | MILEAGE |
| 1 | Correlations | AGE | 1 | -0.106 | -0.668 | -0.325 |
| PETROL\_PLUGIN | -0.106 | 1 | 0.078 | 0.006 |
| EC2 | -0.668 | 0.078 | 1 | -0.014 |
| MILEAGE | -0.325 | 0.006 | -0.014 | 1 |
| EC1.3 | -0.67 | -0.115 | 0.61 | -0.052 |
| Covariances | AGE | 46894.203 | -28461.723 | -87449.837 | -0.688 |
| PETROL\_PLUGIN | -28461.723 | 1551600.243 | 58634.879 | 0.071 |
| EC2 | -87449.837 | 58634.879 | 366002.078 | -0.082 |
| MILEAGE | -0.688 | 0.071 | -0.082 | 9.54E-05 |
| EC1.3 | -91992.583 | -90663.147 | 234112.561 | -0.319 |

Table 10: Coefficients correlations table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | MODEL SUMMARY | |  |  |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | |
| R Square Change | F Change |
| 1 | .955a | 0.913 | 0.909 | 1661.09691 | 0.913 | 230.323 |

Table 11: Model Summary

Table 11 tells us about the r square and adjusted r square value. These values determine if the model is good, average or bad. If the r square value is more than 0.5 then it is considered to be a good model. In our case the value of r square is equal to 0.913 which is greater than 0.5, so it is considered as a good model.

**APPLICATION OF THE MODEL:**

After performing the regression analysis we have derived the equation of the dependent variable.

PRICE = 42294.329 - (0.062 \* MILEAGE) - (6184.484 \* PETROL\_PLUGIN) - (4212.47 \* ENGINE\_CAPACITY\_1.3) - (3303.751 \* ENGINE\_CAPACITY\_2) - (2081.811 \* AGE)

We can now derive the price value of any second hand Mercedes Benz B class vehicle by using this equation.

Consider the following characteristics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mileage | Engine capacity | Engine type | Fuel type | Age |
| 36919 | 1.5 | Automatic | Diesel | 3 |

Table 12: Random sample taken for the application of model

PRICE = 33759.918 POUNDS

**REFRENCES**:

Carsite website:

<https://www.carsite.co.uk/used-car-price-guide/mercedes-benz/b-class>

Autotrader website:

<https://www.autotrader.co.uk/car-details/202110218749657?onesearchad=New&onesearchad=Nearly%20New&onesearchad=Used&year-to=2022&year-from=2017&radius=1500&include-delivery-option=on&sort=relevance&postcode=b129nd&advertising-location=at_cars&make=Mercedes-Benz&model=B%20Class&page=1&percentVehiclePriceDeposit=true>