|  |
| --- |
| DESCRIPTIVE ANALYTICS REPORT  ANALYSIS AND VISUALIZATION OF THE SECOND-HAND CAR MARKET |
| ASTON UNIVERSITY BNM854J – DESCRIPTIVE ANALYTICS  220390110 |



**Table of Contents**

[1.0 SUMMARY 3](#_heading=h.26in1rg)

[2.0 Data Visualization 4](#_heading=h.lnxbz9)

[3.0 Visualisation Methods 4](#_heading=h.35nkun2)

[4.0 Descriptive Statistics](#_heading=h.1ksv4uv) 8

[5.0 Confidence Interval of Average Prices 1](#_heading=h.44sinio)1

[6.0 Hypothesis Testing 1](#_heading=h.2jxsxqh)1

[7.0 Correlation Analysis 1](#_heading=h.z337ya)2

[8.0 Regression Analysis 1](#_heading=h.3j2qqm3)3

[9.0 Residual Analysis 1](#_heading=h.1y810tw)4

[10.0 Conclusion 1](#_heading=h.4i7ojhp)7

[11.0 References 17](#_heading=h.2xcytpi)

**1.0 SUMMARY**

The research aimed to assess the second-hand car buying and selling industry in B27 West Midlands, Birmingham UK, with an emphasis on both average prices and pricing patterns. The analysis was based on Auto Trader's data, which had 300 records, and SPSS's random sample function was used to randomly select 100 records.

In the software versions of Excel and SPSS, descriptive statistics were employed to describe the fundamental properties of the data. Tests such as confidence intervals and hypothesis tests were employed to produce the mean price of the sample. The accuracy of the model was assessed through the use of a parsimonious model and regression analysis, which took into account assumptions like linearity, independence, including uncertainties, and residual normality.

The constructed model was straightforward and explicable, with coefficients providing insight into how a car's price is affected by variables like age, mileage, and make. In the process of evaluating the model's reliability, we utilized various useful measures such as R-squared, adjusted R–squared, and residual standardized error.

Based on car data from 5 years, the study provides information regarding the cars by examining their efficiency during the period and forecasting future trends with greater precision.

**2.0 DATA VISUALIZATION**

The art of data visualization involves visualizing intricate datasets using charts, graphs, or maps. This is a form of presentation known as data analytics. The primary goal is to facilitate the recognition of Information is characterized by designs, patterns, connections and bits of knowledge that convey information and observations to the non-technical as well as business background people. In most of the STEM and social science courses, data visualization plays an important role in taking a decisive step towards information dissemination. It also provides insights into the workings of individuals in their respective fields or departments.

Although purchasers are subject to various considerations when buying a used car, considering things like condition, model, and make, price is often the most significant factor. The initial thought behind taking decision-making process by the customer has been the price and also various other aspects mentioned above. Understanding the aspects that affect prices can be a valuable skill to possess.

**3.0 VISUALIZATION METHODS**

**PRICE DISTRIBUTION OF THE SECONDHAND CAR MARKET**

|  |
| --- |
|  |

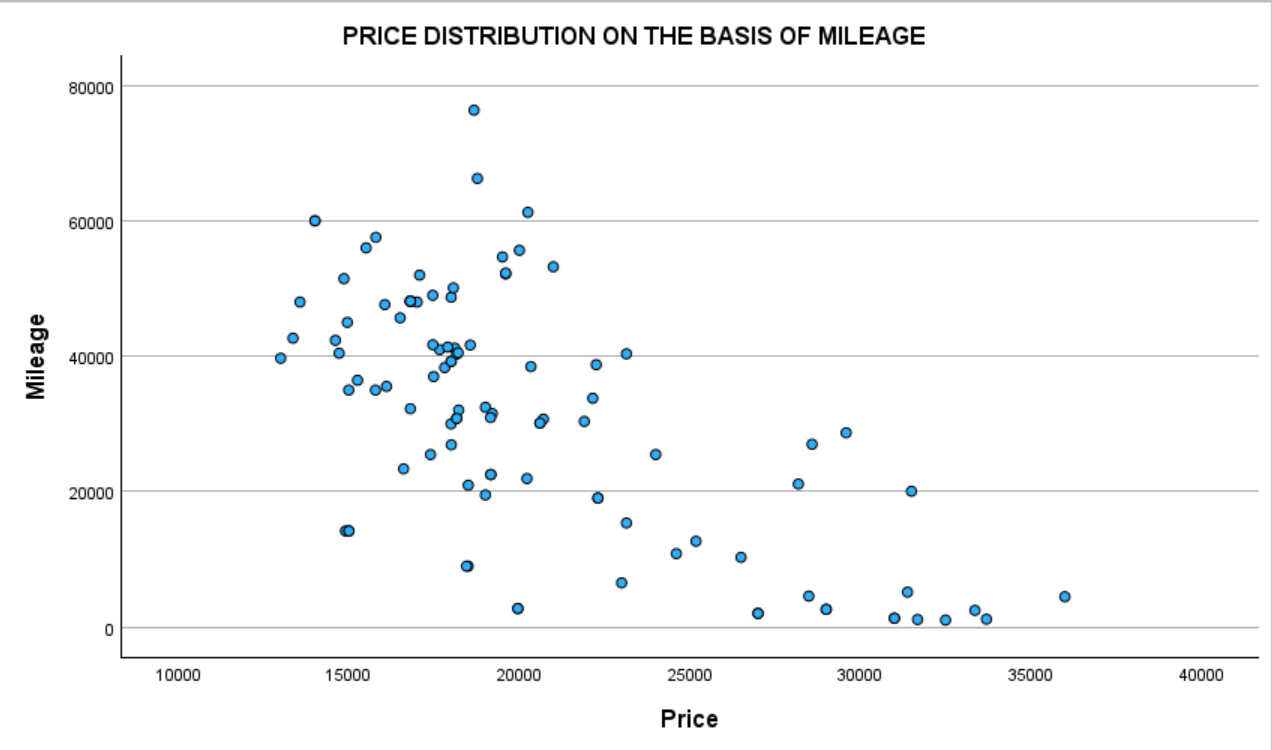
FIGURE 1: HISTOGRAM OF PRICE

(SOURCE: SPSS)

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of bars to the x-axis determines the grouping of data into bars in a histogram.  **Continuity:** The frequency distribution of the price is more easily visible when the histogram is oriented towards the x-axis.  **Figure-Ground Relationship:** Histogram colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The histogram efficiently presents the frequency distribution of price proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Price" on the x-axis and "Frequency" on the y-axis), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the histogram ensures consistency and coherence in the visual presentation. |

|  |
| --- |
| The price distribution for used Mercedes-Benz vehicles is depicted in the histogram. The random sample considered for the analysis of cars is 100. Normal distribution can be done for the sample taken, which indicates that most of the cars are approximately at the high end of the range and few of the vehicles are at the low end of the cost range. The X-axis depicts the cost range of the cars, while the Y-axis indicates the number of cars in the above-considered frequency of price. Most cars cost between $20,000 and $30,000, and the frequency drops significantly when the price drops below this range. |

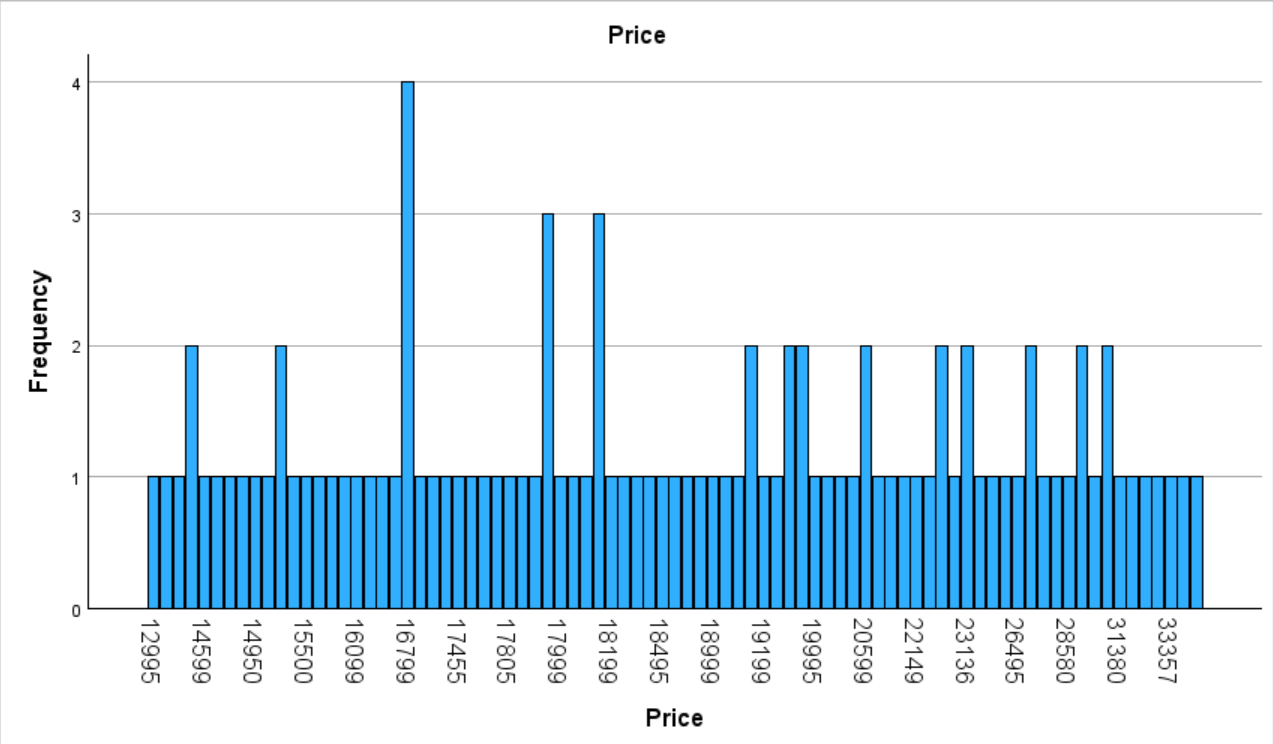
**PRICE DISTRIBUTION BASED ON MILEAGE**



Cost and mileage are depicted on a scatter plot, with X-axis showing the cost and Y-axis depicting the mileage. The scatter plot uncovers a negative relationship between price and mileage. This implies that as the mileage increases, there's a slight and significant dip in cost, but the relationship between the above-mentioned factors is not exceptionally solid. The information that focuses on the plot is spread broadly, demonstrating that there's a considerable amount of change and movement possible within the information. As the relationship coefficient is dipping low to zero, these two streams are negatively influencing each other, proposing that the factorial credibility is not providing favourable circumstances.

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of dots to the x-axis determines the grouping of data into scattered information in a scatterplot.  **Continuity:** The distribution of the price is more easily visible when the price is oriented towards the x-axis.  **Figure-Ground Relationship:** Scattered dot’s colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The Scatterplot efficiently presents the distribution of price proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Price" on the x-axis and "Mileage" on the y-axis), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the dots ensures consistency and coherence in the visual presentation. |

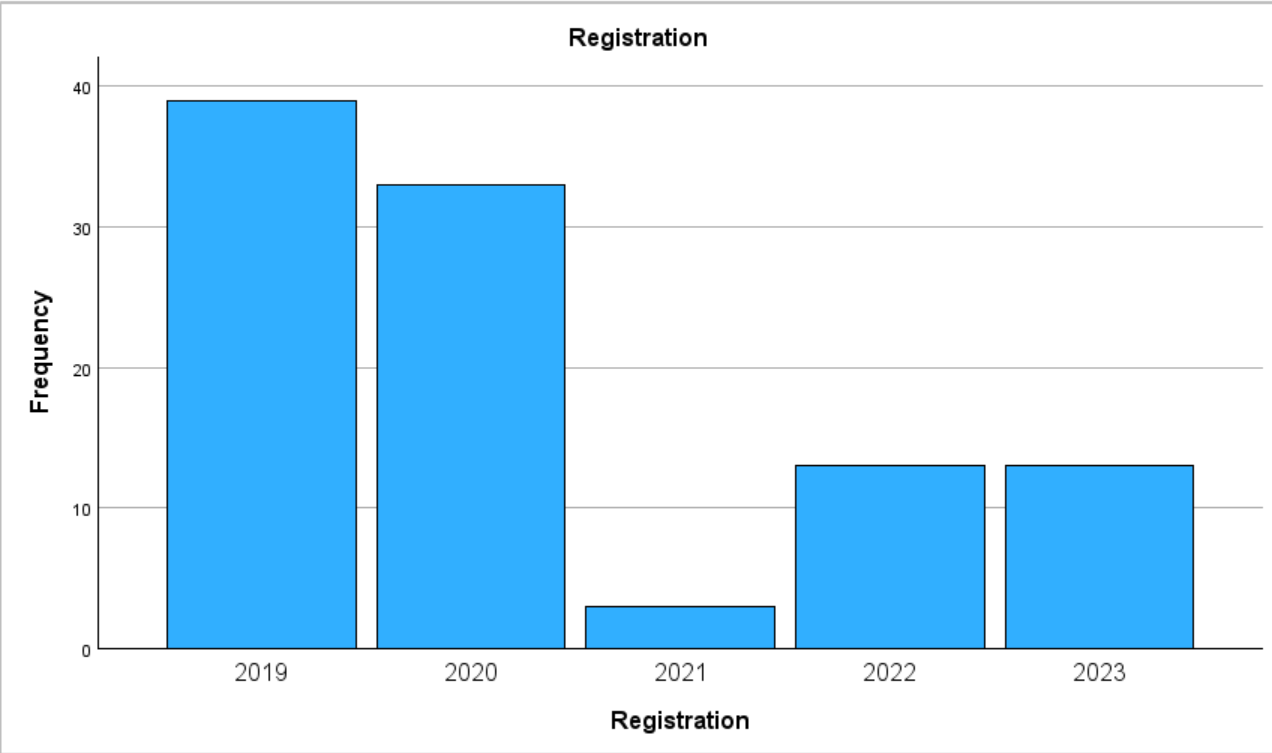
**FREQUENCY DISTRIBUTION BASED ON PRICE**



Car prices between $12,995 and $35,995 are distributed in the data set. An SPSS bar chart that shows the occurrence of each price group is displayed, with the range being included. The information is clear. The chart indicates the number of price fluctuations, with some being more significant than others. The distribution of car prices in the data facilitates a better understanding of market patterns and trends.

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of bars to the x-axis determines the grouping of data into scattered information in a bar plot.  **Continuity:** The distribution of the price is more easily visible when the price is oriented towards the x-axis.  **Figure-Ground Relationship:** The Bar’s colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The bar plot efficiently presents the frequency distribution of price proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Price" on the x-axis and “Frequency" on the y-axis), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the bars ensures consistency and coherence in the visual presentation. |

**FREQUENCY DISTRIBUTION BASED ON THE REGISTRATION**



Car registrations between 2019 and 2023 are distributed in the data set. An SPSS bar chart that shows the occurrence of each year is displayed, with the range being included. The information is precise. The chart indicates the number of cars bought according to the years are different, with some being more significant than others. The distribution of cars in the data facilitates a better understanding of market patterns and trends. Based on the year, cars with different fuel types such as Petrol Plug-in hybrid and Petrol hybrid can be observed more frequently and rose to a good level.

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of bars to the x-axis determines the grouping of data into scattered information in a bar plot.  **Continuity:** The distribution of the registration is more easily visible when the registration is oriented towards the x-axis.  **Figure-Ground Relationship:** The Bar’s colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The bar plot efficiently presents the frequency distribution of registration proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Registration" on the x-axis and “Frequency" on the y-axis), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the histograms ensures consistency and coherence in the visual presentation. |

**4.0 DESCRIPTIVE STATISTICS**

In statistical analysis of samples, variables are used to classify the data based on the characteristics of the car being studied. Factors like price, mileage, capacity, power, and year of manufacture can be included in these variables. We can determine their relationship by examining the data through these variables.

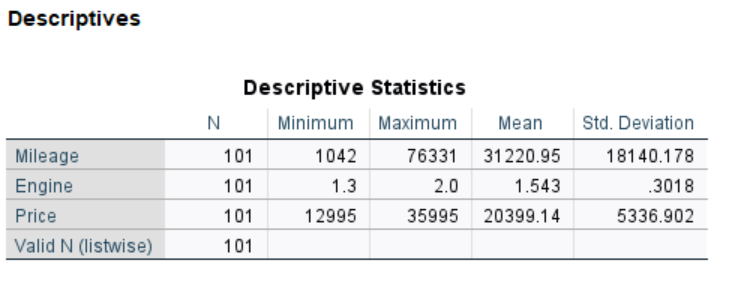
**DESCRIPTIVE STATISTICS FOR NUMERICAL VARIABLES:**

**Descriptive stats for price**

A screenshot of a phone

Description automatically generated

TABLE 1



The information set containing 100 samples of cars are taken for the analysis, with the costs estimated around £12,995 for the cheaper one and the maximum price at £35,995. After performing the statistical procedures, the cars fetched £20,399.14 as the mean, and with a variance of £28,477,815.81004. The estimated miles travelled by vehicles is approximately 31,220.95, with the minimum travelled count being 1,042 miles and 76331 being the highest and most used vehicle in terms of miles. The major difference in mileage comes about by seeing the tall standard deviation of 18140.178. The car engine size yield ranges from 1.3 to 2.0, with an average of 1.5. In general, the information set gives profitable experiences by showcasing the costs, usage, and motor size of 100 cars, exhibiting the variety and separation of frequencies.

**MEAN GEARBOX TO COST**

**A screenshot of a graph

Description automatically generated**

TABLE 3

The Automatic and Manual gearbox types are represented using the binary and Boolean terms such as 0 and 1. The average costs based on the report suggest the interest shown by the people towards the Automatic gear types is more when compared to the Manual types even though the price of the latter is lesser than the former.

**MEAN MILEAGE AND FUEL TYPE**

A screenshot of a graph

Description automatically generated

In the data 0 represents Diesel, 1 represents the Petrol, 2 represents the Petrol Plug-in Hybrid and 3 represents the Petrol Hybrid. We can see that the mileage of Petrol Hybrid versions is lesser when compared with the other types and also Petrol Plug-in hybrid type has the highest mileage followed by Diesel and Petrol counterparts.

**MEAN ENGINE TO FUEL TYPE**

A screenshot of a graph

Description automatically generated

In the data 0 represents Diesel, 1 represents Petrol, 2 represents the Petrol Plug-in Hybrid and 3 represents the Petrol Hybrid. We can see that the Engine Size of the Petrol Plug-in Hybrid versions is smaller when compared with the other types and also Diesel type has the highest followed by the Petrol Hybrid Fuel type. Astonishingly, Petrol’s vehicles motor capacity but we cannot completely stick to the conclusion as the sample size of Petrol vehicles is more than half of the proportion.

**MEAN MILEAGE TO GEARBOX**

A screenshot of a graph

Description automatically generated

The 0 and 1 represents Automatic and Manual modes of gearbox. The estimate from the report suggests that the sample has more Automatic geared cars when compared to the Manual and the mileage of the manual type is more compared to that of the automatic.

**5.0 CONFIDENCE INTERVAL OF AVERAGE PRICES**

After analysing the sample with a confidence level of 95%, the below-derived outputs can be obtained using the cost of the cars as the primary focus.

A screenshot of a data

Description automatically generated

The table outlines the confidence intervals for the mean price of the Mercedes A-Class car with in the second-hand availability range and dealerships. Assuming a normal distribution with 95% confidence, the sample data and the mean of the interval are £20,399.14. The estimated confidence interval is between £12,995 and £35,995.

**6.0 HYPOTHESIS TESTING**

A one-sample t-test was conducted to test our hypothesis with 95% confidence. The test was considered appropriate due to its comparison of the mean of a single sample with the mean given by a single population value.

|  |
| --- |
|  |

The calculated t-value of the one-sample t-test was -0.171, compared to the critical values ​​of +/- 1.96 for the 0.95 confidence interval level. This range of the output t-value makes it impossible to rule out the H0 (Null-Hypothesis) thus rejecting the HA (Alternate Hypothesis). The binomial tow-tailed p-value gathered from the analysis is 0.865, which indicates that the test value can be in the range of 0.95 confidence level. Based on the analytical recollection, a conclusion can be obtained which suggests that the average cost of the car can be calculated using the sample mean.

**7.0 CORRELATION ANALYSIS**

The Pearson correlation coefficient could be a numbered range that can assess the relational capacity of both components within the limits of +/-1. Correlation is deemed null and void if the value turns out to be 0. Furthermore, the matrix of relatability cements the situation to acknowledge and identify the dependence of one variable on the other.

|  |
| --- |
| A table with numbers and text |

A correlation between factors, variables, and factors is evident from the table provided. Positive correlation implies the subsequent and stable growth and upward trend of both variables based on the other variable. Conversely, a negative correlation indicates that the increasing value of one variable implies a decrease in the values of the other variable.

**8.0 REGRESSION ANALYSIS**

The best model was determined by using multivariate regression analysis, in which the evaluation of outcome of the dependent (Y) variable which can be obtained based on its correlation type with the independent (X) variables which can be one or many based on various situations. We employed and performed numerous analyses to get significant desirable price output and to eliminate ambiguity in variables. The final model includes the variables that are important to correctly predict the price by keeping the significance in mind.

|  |
| --- |
| A close-up of a number |

Cost of the vehicle depends upon multiple reasons. Among them if considered any two, the most significant variability is due to the Registration (Year) and miles car travelled. The sum of squares of both regression and residual analyses models explains and predicts the difference among the costs of cars. It's clearly shown that the fuel type and gearbox classification are also the major influencers in finalizing car prices.

|  |
| --- |
| A table with numbers and a price |

6

|  |  |
| --- | --- |
| **REGRESSION SUMMARY** | A summative chart demonstrates that the adjusted R2 value is 0.762. 76.2% probability by which the exact and ultimate solution of the effective costs which are Y variables can be predicted by this model. Analysis of significance value in the table concludes that the rest of the variables must have a value less than a 0.95 confidence level interval. Thus, in multiple linear regression analysis, the null hypothesis of a significant relationship between variables can be rejected. |

# 9.0 RESIDUAL ANALYSIS

# The following graphs represent the residual analysis of a parsimonious model

# 9.0 Residual Analysis

|  |
| --- |
| A graph of a normal distribution  Description automatically generated |

|  |  |
| --- | --- |
| **Graph Explanation** | The data provided includes frequency counts for various values ​​of an attribute called "range". The data format shows the interval distribution, with each unique interval value listed and its corresponding frequency count. From a total of 300 valid population, looking at the frequencies, we can see a wide range of values, with some values ​​being higher than others. Each threshold value corresponds to a horizontal bar on a bar graph, created using SPSS, and the height of each bar is calculated as a frequency count. This view shows exactly the distribution order of the limit values ​​and the frequency of each value. |

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of bars to the x-axis determines the grouping of data into scattered information in a histogram plot.  **Continuity:** The distribution of the Regression Standardized Residual is more easily visible when the two graphs are compared with a histogram and two-tailed curve to establish the regression with a good significance value towards the x-axis.  **Figure-Ground Relationship:** The Bar’s colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The bar plot efficiently presents the frequency distribution of Regression Standardized Residual proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Regression Standardized Residual" on the x-axis and “Frequency" on the y-axis), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the histograms ensures consistency and coherence in the visual presentation. |

|  |
| --- |
| A graph of a normal graph  Description automatically generated with medium confidence |

|  |  |
| --- | --- |
| Gestalt Principles: | **Proximity:** The proximity of dots from the x-axis determines the grouping of data into scattered information in a normal probability plot.  **Continuity:** The distribution of the residual statistics is more easily visible when the residuals are oriented with slope and constant included.  **Figure-Ground Relationship:** The dots and line colours create a connection between images and the background, which helps identify data in any area of the workspace. |
| Tufte’s Theories: | **Minimised Chart Junk:** The layouts are clean and minimal, with just enough detail to accurately display the data.  **Maximised Data Density:** The P-P plot efficiently presents the cumulative probability distribution with price as the dependent variable with proximity in a compact and informative manner. |
| IBCS Guidelines: | **Clear Labelling:** The axes are clearly labelled with the corresponding variables and axis titles ("Cumulative Probabilities on both the axes” (Observed and Expected)), ensuring clarity.  **Consistent Use of Colour:** The use of colour for the dots ensures consistency and coherence in the visual presentation. |

|  |  |
| --- | --- |
| **RESIDUAL ANALYSIS SUMMARY** | Residual analysis plots may showcase subtle elements that correspond to the measurable residual values of the regression model. The predictions of the model include the observations from the past since the mean residual value is zero, most of the estimates are usually correct. Due to the fact that the standard deviation of residuals is not exceptionally high, the estimates of car prices are typically modified. The numbers in the residuals differ significantly, ranging from negative to positive. The following items may offer insights into the accuracy of the expectations. |

**10.0 CONCLUSION**

# An in-depth exploration revealed that car prices are subject to various

# factors. Estimated costs, mileage, motor type, fuel classification, gearbox classification (fuse), body classification, and colour are some of the factors that determine a car’s cost. In addition, examining relationships highlights that these factors are undoubtedly correlated. It is possible to calculate the amount of a car that will be charged by using the zone and number of cars, while also asking for information on employment history. The knowledge of these things in order is essential for accurately predicting and comprehending the movement of car markets.

# 11.0 REFERENCES

Autotrader.co.uk. (2019). *Auto Trader UK - Find New & Used Cars for Sale*. [online] Available at: https://www.autotrader.co.uk/.