Siddhesh Laxmikant Panchal

220363365@aston.ac.uk

Student ID- 220363365

2nd Hand Car Market Report

Descriptive Statistics

**TABLE OF CONTENTS**

**INTRODUCTION2**

Objective2

Executive Summary2

Sources2

Limitations2

Procedure of sampling3

Duration of 5 years3

Software used3

**DATA VISUALIZATION3**

Scatter Plot3

Histogram4

Clustered Bar Graph5

Bar Graph6

**DISCRIPTIVE STATISTICS6**

**CONFIDENCE INTERVAL ANALYSIS7**

**HYPOTHESIS TESTING**8

**CORELATION STATISTICS9**

**REGRESSION MODEL9**

**RESIDUAL ANALYSIS10**

**STATISTICAL MODEL12**

**CONCLUSION13**

**APPENDIX13**

**INTRODUCTION**

**OBJECTIVE**

The goal of the report is to do a survey of the market to study the second-hand car market in the United Kingdom for a Volkswagen Tiguan model. The study's purpose is to determine the most important characteristics that influence the price of a used car and to develop a statistical model that can be used to anticipate the market value of a car based on miles driven, engine size, and other optional features. Apart from examining the automobile market in a certain post code, it is also required to compare it to the total car market for Volkswagen Tiguan models in the United Kingdom.

**EXECUTIVE SUMMARY**

The automobile market in United Kingdom for the model Tiguan of Volkswagen are being analyzed after gathering reliable data from the website [www.autotrader.co.uk](http://www.autotrader.co.uk) for the post code M24 WU. The data is from the years 2019-2023, and it includes the car's selling price, mileage, engine size, and fuel type. Thus, all these components are critical for conducting analysis. The report uses various visualizations, such as scatterplots, bar graph, and histograms, to describe the data, summarize descriptive statistics tables, and discuss the information obtained as well as various statistical analyses such as residual analysis, correlation, regression analysis, hypothesis testing, and so on. All of this is considered in order to generate a reliable overview of the used automobile market in United Kingdom.

**SOURCES**

The automobile market in United Kingdom for the model Tiguan of Volkswagen are being analyzed after gathering reliable data from the website [www.autotrader.co.uk](http://www.autotrader.co.uk)

**LIMITATIONS**

The research's limitations start from its 5-year timeframe. Also, the sample size obtained is 80-100, which does not represent the full population of used cars in that post code, which can undermine credibility. Furthermore, the study only covers a few criteria; other key elements that can affect the pricing are not considered.

**PROCEDURE OF SAMPLING**

Stratified Random sampling is utilized as the sampling method to create a reliable report on the analysis of the secondhand car market in the United Kingdom.

**DURATION OF 5 YEARS:**

By narrowing the age range, the study may concentrate on more current models, which are expected to have more consistent prices and can provide more accurate insights into the factors that influence the value of used automobiles on the market.

**SOFTWARE USED:**

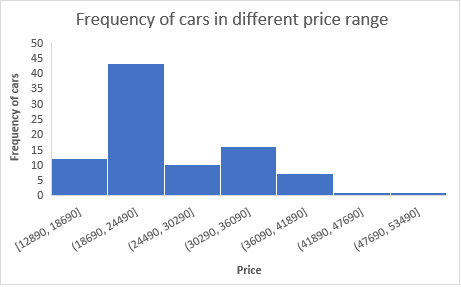
Excel and SPSS were the software which were used to derive the analysis for the secondhand car market.

**DATA VISUALIZATION**

1. Visualizing the relationship between Price and Miles through Scatterplot.

Chart, scatter chart

Description automatically generated This scatter plot shows that Miles is the dependent variable in connection to Price, and that as Miles drops, Price rises. **Tufte's principle of graphical integrity** concludes that the graph is proportional because the y- and x-axis both contain zero. The graph is clear because it indicates which axis is which. It is also not completely standardized because the units are absent, which might lead to incorrect data interpretation. The labelling is partially visible because the graph is accurately labelled but lacks context, such as whether the price is in thousands or millions of dollars. The **graphical excellence** is achieved by reducing data ink by removing gridlines. Outliers are identified using a unique color, according to **the Gestalt Principles** of Similarity. According to **IBCS** the above graph has correct variables chosen. The inference of the data is visually represented from the graph. Accuracy of the data and identification of outliers is shown.

1. Visualizing frequency of cars for a various price range.

Chart, histogram

Description automatically generated This histogram depicts the quantity of cars available in various price ranges. The graphs shows that the price range 18690-24490 has the most number of cars. **Tufte's principle of graphical** integrity states that the graph is not truly proportional because the y-axis does not include zero. It is also not completely standardised due to the lack of units which misleads the data. The labelling is partially visible because the graph is accurately labelled but lacks context, such as whether the price is in thousands or millions of dollars. The **graphical excellence** is achieved by reducing data ink by removing gridlines. A separate colour can be used to denote different price ranges according to the **Gestalt Principles** of Similarity. According to **IBCS** the above graph has correct variables chosen. The inference of the data is visually represented from the graph. Accuracy of the data and identification is shown rightly.

1. Visualizing Sum of price for different types of fuel type according to the years.

Chart, bar chart

Description automatically generated This clustered bar graph illustrates the aggregate of prices for several fuel kinds over time. It consist of sum of price on the y-axis and years on the x-axis and the feul type is our chart legend. This graph shows that most of the cars are petrol and hybrid cars arent present in the year 2019 and 2020 but are present in the year 2021 and 2022. Diesel can be seen the most number in the year 2022. According to **Tufte's graphical integrity principle**, the graph is not truly proportional because the y-axis does not include zero. This graph has a good clarity because the axis are labelled properly and the title is there. Due to a shortage of units which missleads the data, it is also not totally standardised, because the graph is not appropriately labelled and lacks context, such as whether the price is in thousands or millions of dollars, the labelling is only partially apparent. The **graphical excellence** is achieved by reducing data ink by removing gridlines.. According to the **Gestalt Principles** of Design, different fuel types are denoted by a different colour. According to **IBCS** the above graph has not presented the variables correctly. The inference of the data is visually represented from the graph.

1. Visualizing count of different models in Volkswagen.

This bar graph depicts the count of cars available in different models of Volkswagen. This graph shows that Tiguan model of Volkswagen has the most frequency when compared to other models. **Tufte's principle of graphical integrity** states that the graph is not truly proportional because the y-axis does not include zero. It is also standardised as required units of the graphs are present. The graph is accurately labelled and and has clearity in what the graph is Chart, bar chart

Description automatically generatedpresenting. The **graphical excellence** is achieved by reducing data ink by removing gridlines. A separate colour has been used to the differentiate the car model having the highest number of counts according to the **Gestalt Principles** of Similarity. According to **IBCS** the above graph has correct variables chosen. The inference of the data is visually represented from the graph. Accuracy of the data.

**DESCRIPTIVE STATISTICS**

Table

Description automatically generated

This is a statistics summary table for three variables - Price, Mileage, and Engine Size - for a sample of 90 used cars.

A standard deviation shows how far away a particular value is from the mean.

The mean price of the 90 cars is £25,716.06, with a standard deviation of £7,487.67. The standard deviation is 25,170.90 miles, with a mean mileage of 30,581 miles.

A standard error shows how much difference is there between the sample population mean and entire population mean.

If there were a greater number of samples there would be less standard error as there is a error of 789.28 for the price column and 2653.25 for the milage column, but in comparison to this there is a very minute standard error for the engine size column which is 0.02

Table

Description automatically generated

The above descriptive analysis is done for pricing but divided by fuel type. This statistical analysis shows the average price of diesel, hybrid, and Petrol which is 23785.72£, 38048.43£, 25234.57£, 25716.06£ respectively. Which shows that Hybrid cars have the highest average price.

**CONFIDENCE INTERVAL ANALYSIS**

A confidence interval analysis shows what will be the range in which the average mean price will fall.

Timeline

Description automatically generated

Considering that prices of the secondhand cars are normally distributed. The average price range of the secondhand cars will fall between 24147.79£ to 27284.32£ at 95% Confidence Interval

**HYPOTHESIS TESTING**

Hypothesis testing is a statistical analysis in which we assume two occasions between two population parameters and find the correlation between them by rejecting one of the two occasions.

|  |  |  |
| --- | --- | --- |
| One Sample Test | | |
|  | M24 WU | National |
| Mean | 25716.06 | 26563.30 |
| P value |  | <0.001 |

The p value is <0.001.

A picture containing timeline

Description automatically generated

As we can see above the p-value is extremely small (p < 0.001) which is less than the significance level (𝛼=0.05), we can conclude that there is a significant difference between the mean price of the secondhand car in our area code and mean price of the secondhand car in the entire United Kingdom. But we can’t rely on this completely as the sample size of the assumption was only 90 cars.

**CORRELATION STATISTICS**

Correlation statistics show the relationship between two variables 1 being highly related, -1 being negatively related and 0 mean no relation.

Table

Description automatically generated

The above table displays the correlation coefficients and statistical significance levels for four variables: price, year, miles, and engine size.

In the table above, we can see that there is a positive Pearson correlation between price and year, which is 0.800, indicating that as the year of manufacture rises, so does the price of the car. The price and mileage relationship are inversely related by -0.665; as the mileage increases, the price of the car decreases.

**REGRESSION ANALYSIS**

Table

Description automatically generatedRegression analysis is a statistical analysis which estimates what is the relationship between a dependent and one or more independent variable.

Table

Description automatically generated

Table

Description automatically generated

This is our Parsimonious Model with the model accuracy of 75.1%. as all the variables are significant.

With the above table we can calculate the value of our regression model =

(Constant) + (Miles \* -0.97) + (Year\_21 \* 7885.096) + (Year \_22 \*10083.663) + (Hybrid \* 4681.842)

**RESIDUAL ANALYSIS**

Residual analysis is used to calculate how much error there is between the observed value and predicted value.

Chart, scatter chart

Description automatically generated

Assumption 1: In the above scatter plot, it contains roughly the same number of residuals above and below zero, thus the assumption holds.

Assumption 2:

In the above scatter plot the residuals aren’t scattered in a particular pattern they are scattered independently. Hence the assumption Holds.

Assumption 3:

The scattered residuals don’t show any pattern of increasing or decreasing values its scattered randomly. Thus, the assumption holds.

Chart, histogram

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Assumption 4:

In the above histogram the residuals are normally distributed forming a bell-shaped curve and in the above P-P Plot the residuals are distributed closer to the line of perfect normal distribution. Hence the assumption holds.

Table

Description automatically generated

Assumption 5:

In the above correlation matrix, all the independent variables have their Pearson Correlation values less than 0.7. Hence, the assumption 5 holds.

**STATISTICAL MODEL**

Statistical Model is a mathematical equation which is formulated according to the observed data.

Now to Determine the value of Volkswagen Tiguan model of the manufacture year 2021 which has milage of 11407 and having engine size 2 and fuel type as Diesel with the below derived statistical model from regression analysis.

(Constant) + (Miles \* -0.97) + (Year\_21 \* 7885.096) + (Year \_22 \*10083.663) + (Hybrid \* 4681.842)

The value of the car is £21925.687.

Hence we have calculated the statistical model of Volkswagen Tiguan, which can be used for prediction of other models given that there are all the parameters such as engine size, year of manufacture, fuel type and milage present as they are the important variables for the prediction of price.

**CONCLUSION:**

In the preceding research, we began by determining the association between the car's price, miles, engine size, year of production, and fuel type. Following that, considering the limitations we learned how other elements influence the price, which enabled us to create various types of graphs. Following that, we reported what the car's average price range should be at 95% confidence interval, and then we compared the average price of our post code to the average price of the whole United Kingdom with the help of hypothesis testing. Following that, we developed a regression model with 75.1% accuracy, on which we performed residual analysis and passed all the five assumption tests. The regression model provided us with a statistical model that will assist us in predicting what the price of the car should be based on the year of production, engine size, fuel type, and miles travelled, as demonstrated in the preceding research.

**APPENDIX**

**Confidence Interval**

We chose a 95% confidence interval which means that we have a 5% chance of being wrong. To present the confidence interval we use One sample Test as we have one data sample which we want to consider in this case which is price, and we’d like to know at 95 confidence interval what is the price range in which our average secondhand car price lies.

The one sample test takes in the values for the variable price and calculates its upper and lower bound values with respect to the mean and standard deviation it gives us a price range . In our case the price range is 24147.79 to 27284.32 pounds sterling.

**Hypothesis Testing**

We want to check the average’s a difference or no between Average national price of Tiguan model of Volkswagen and average price of our post code.

For this we consider our null hypothesis as there is no difference between the mean price and consider our alternate hypothesis as there is difference between the mean price.

We take a significance level of 0.05.

For this we use One sample test as it checks whether there is a statistical difference between the mean price of a particular population and our hypothesized value.

In One Sample test the test variables mean is considered against a test value and in return it gives a probability value which we call as p-value and that we compare with the significance level . If the P-value is less than the significance level, we reject the null hypothesis and if its greater than the significance we fail to reject the null hypothesis.

Table

Description automatically generated**Regression Model**

The Model summary of our regression model says that there is not much of a difference between R Square and Adjusted R Square, so we take Adjusted R square into consideration which depicts that our model is 74.5% Accurate.

Table

Description automatically generated

However, the significance level of our Engine\_2, Year\_20 and Diesel is greater than 0.05 in our coefficient table which explains that this model is not a parsimonious model. Hence, we must run our linear regression model once again but this time removing them.

So, we get a model summary of our regression model which says that our model is 75.1% Accurate.

According to the ANOVA results, the independent factors (Miles, Year and Fuel Type) have a significant effect on the dependent variable (price) since the significant value (<0.001) is less than 0.05.

**STATISTICAL MODEL**

(Constant) + (Miles \* -0.09) + (Engine\_2\* 7844.128) + (Year\_21 \* 7062.239) + (Year \_22 \* 9987.904) + (Diesel \* -7513.380) + (Hybrid \* 5291.63)

Let’s calculate the value if the car by substituting the values in the above regression model.

25105.381 + (11407 \* -0.97) + (1 \* 7885.096) + (0 \* 10083.663) + (0 \* 4681.842)

= 25105.381 - 11064.79 + 7885.096 + 0 + 0

= 21925.687

The value of the car is £21925.687.

The actual value is £23568.

Hence the statistical model is a **good fit** as there is not much of a difference between the actual price and predicted price.