12/16/2022

DESCRIPTIVE ANALYTICS COURSEWORK

Analysis of UK second-hand car market

ANANDU KARUNAKARAN
220241328

TABLE OF CONTENTS

Table of Contents

INTRODUCTION

This is a business report done on behalf of a market research company that is acting on behalf of Car4all, on the second-hand car market in the UK. The source of data collected is the website motors.co.uk. The car chosen for the analysis is Ford fiesta and data collected is from the post code B297PU Birmingham within a 20-mile radius. Ford fiesta is one of the most common cars in the UK market. Web scrapping was used to collect the data from the website and one of the main challenges faced when collecting the data was that some information got missed during the scrapping process and some information was not available on the website. After collecting the data, it had to be cleaned properly to obtain the proper data ready for further analysis.

SAMPLING METHODOLOGY

The sampling method used for sampling the population is stratified random sampling. Stratified sampling was chosen for sampling the data because through this sampling method we can obtain almost same proportion of all variables in the population which enables us to obtain a representative of the population. Stratified sampling ensures that all subgroups of the population are included in the population. After sampling, there was 100 observations in the data with some outliers detected using z score value of the respective variable and was removed. After outlier treatment, the sample size reduced to 95 observations.

EXECUTIVE SUMMARY

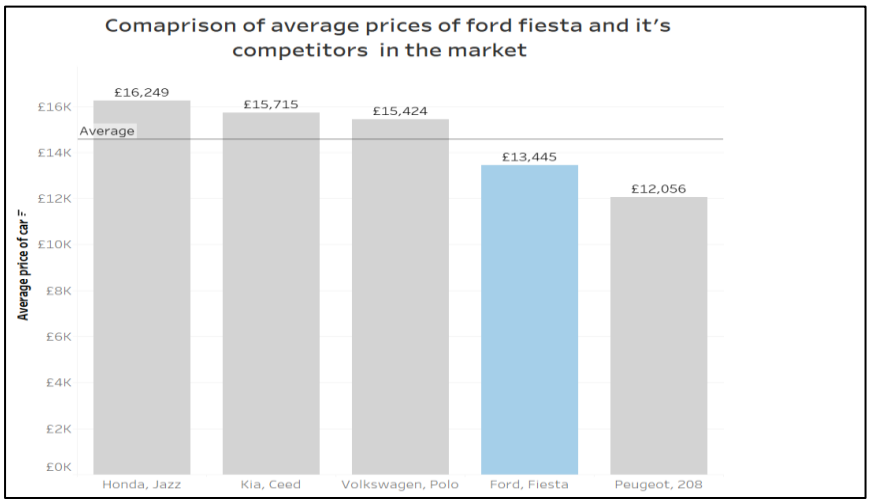
We have analysed the second-hand car market by analysing the relationship between price of the car and other characteristic that affects the price of car like its miles run, performance (BHP), age of the car and its gear box. Its clear from the report that most cars in the market had a manual gear box and had a higher variation in price compared to other modes. The average price of Semi-Automatic cars is comparatively higher than automatic or manual cars. Price of car had a negative linear relationship with age.

From hypothesis testing, it was observed that the average price of cars in our sample is different from the average price of ford fiesta cars in the entire UK market.

From regression analysis, we built the most parsimonious model to predict the price of cars using other variables such as miles run by the car, its performance, age, and which type of gear box the car had. After completing the regression analysis by checking the model's goodness of fit, we obtained a model where the independent variables could explain the 80.3 % variability in price. As the adjusted R square which is the value that is used to test the goodness of fit of the model is very high, it can be concluded that the model is very good and can predict the price of the car with good accuracy. The model's adequacy was confirmed by checking if the model satisfies the four assumptions

DATA VISUALISATION

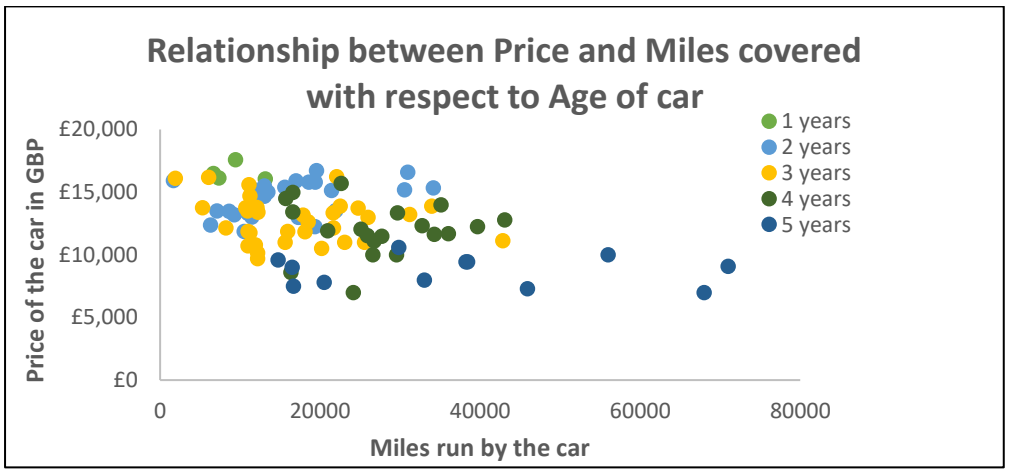
Graph 1



The above graph shows the average price of 4 different competitors of ford fiesta in the B24 region of Birmingham. It is clear from the graph that, ford fiesta is the second cheapest car out of the five, just above Peugeot 208. Average price of ford fiesta is lower than the overall average price of the five cars. Popularity of ford fiesta in the UK market can be due to the fact that its average price is comparatively lower than its main competitors.

Gestalt's principle of proximity is being followed here. That is, the average price of each car is given on top of each bar that**. Tufte's principle of graphical integrity** is followed by the graph as there **is proper labelling, least amount of ink used to show relevant findings**. The graph is **multivariate** as it compares two variable price and different models of the market. The graph follows **proportionality** as all the measures start from zero and proper comparison of each bar is achieved, and appropriate number of **dimensions** is used to plot the graph. Different colour is given to the bar of ford fiesta as the main attention in the graph is given to ford fiesta.

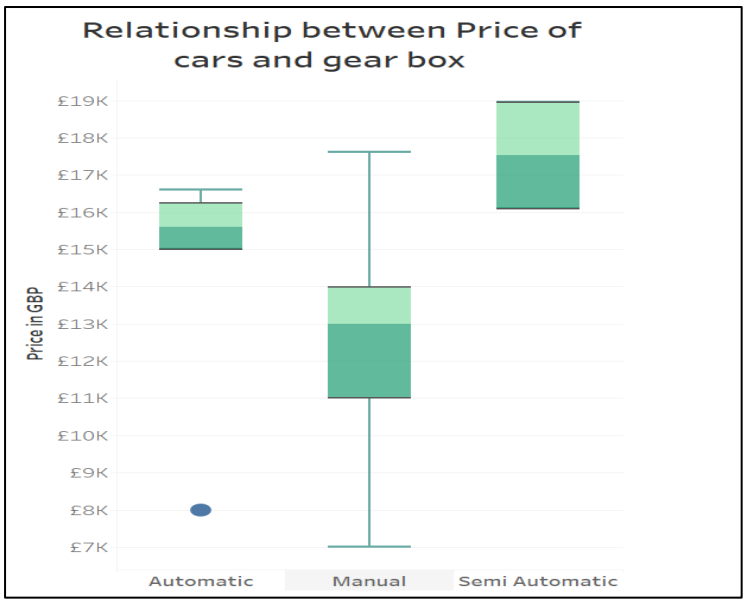
Graph 2



From the above graph we can see that with the increase in miles run by the car there is decrease in price of the car. Another observation from the scatter plot is that as car gets older, its price tends to decrease. As miles run by the car increases and it gets old, its price will start to decrease.

The graph satisfies **Gestalt's principle of similarity** as colour is used to identify the age of each car in the graph. The graph is developed with proper labelling of axis and titles and gridlines are removed. Appropriate number of **dimensions** are used, and the graph is **multivariate** with three variables age of the car, price and miles run by the car. **Scales** of the graph are properly **standardized**, and appropriate graph is used to find relationship between two numeric variables.

Graph 3



From the box whisker plot between Price of car and gear box, it can be observed that Semi-Automatic cars are in the highest price range compared to Manual and Automatic cars.

Manual cars are in high concentration in the market and have the highest variation in price. This can be due to the fact that Automatic and Semi-Automatic cars might be relatively new. Manual cars have lowest median price of around £13000. It can be observed that there is an outlier for the plot of automatic cars which can be due various reasons. It could either be a defective model or could be comparatively older than the other automatic cars.

The graph follows **Gestalts principle of similarity**. Graph is properly **labelled with title and axis.** The axis does not start from zero in this case because it would make the graph difficult to interpret as price of cars starts from around £7000

3.DESCRIPTIVE STATISTICS

Descriptive statistics gives the basic summary of the data

SUMMARY OF CONTINUOUS VARIABLES

	price	Miles_run	BHP
count	96.000000	96.000000	96.000000
mean	12789.791667	21425.656250	106.572917
std	2571.416767	13055.003283	19.466433
min	6995.000000	1666.000000	82.000000
25%	11079.500000	12065.500000	95.000000
50%	13020.500000	18307.500000	100.000000
75%	14993.000000	26987.750000	125.000000
max	18956.000000	71000.000000	200.000000

- From the summary table we can see that the average price of ford fiesta in the locality is £12789.8.
- Min £6995.00 indicates that the cheapest car available in the market is priced at £6995.00
- Average miles run by a car in the locality is 21425.65 miles.
- Arounf 50 % of cars had a performance or horsepower of 100 BHP.

SUMMARY OF CATEGORICAL VARIABLES

	gear_box	AGE
count	96	96
unique	3	5
top	Manual	3
freq	89	33

- From the above summary table, we can see that top Gear box is manual which indicates that most vehicles have a manual gear box. 89 vehicles out of the total 96 vehicles are manual vehicles
- Most vehicles in the locality are 3 years old. Out of the total 96 cars, 33 are 3 years old.

SUMMARY TABLE OF PRICE VS GEAR BOX

PRICE				
GEAR BOX	**Mean**	**Standard deiviation**	**Max**	**Min**
Automatic	14288.2	3571.145923	16599	7995
Manual	12599.14607	2416.118391	17599	6995
Semi Automatic	17527.5	2020.204074	18956	16099

The above table shows the summary table of price variable divided on the basis of gear box

- The average price of Semi-Automatic cars is considerably higher than the other two types of gear boxes.
- Standard deviation is high for automatic cars which indicates that price variation is high in the automatic cars.

CONFIDENCE INTERVAL

Confidence interval is used to determine the confidence limit with in which the population average price of sample they can lie with a pre-determined level of confidence

From calculating the confidence interval, we can say with 95% confidence that the population average price lies between **£13304.18 and £12275.40**

$$CI = \bar{x} \pm z_{a/2} * (\sigma / \sqrt{n})$$

$\bar{x}$ = 12789.79 (sample mean)

σ = 2571.417 (sample standard deviation)

n= 96 (sample size)

HYPOTHESIS TESTING

One sample t-test: To check whether the average price of car in the sample represents the average price of the car in the entire UK market.

We need the average price of ford fiesta for the entire UK market which can be obtained from motors.ac.uk (https://www.motors.co.uk/car-price-guide/ford/fiesta/).Average price

for population is obtained by calculating the average of average prices of cars with age ranging from 1 to 5.

The obtained average price for UK market is £13737

The steps are as follows

Null hypothesis:
Ho: $\mu = 13737$

Alternative hypothesis:
Ha: $\mu \neq 13737$

– The significance level chosen for this test is **5% (0.05)** and the critical value za/2 = **1.96**

From the sample,

The average price of car $\bar{x}$ =£12789.79

Calculating the test statistics

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
price	96	12789.79	2571.417	262.444

One-Sample Test

Test Value = 13737

	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	-3.609	95	<.001	<.001	-947.208	-1468.23	-426.19

From looking at the above test table we can see that the p value is less than 0.05 and hence we reject the null hypothesis that average price is £13737

- Since we rejected the null hypothesis**, we cannot say that average price of car is £13737 hence the average price of ford fiesta in the postcode B297PU is different from the average price in the entire country.**

CORRELATION ANALYSIS

Correlation is used to check the linear relationship between numerical variables. It ranges between -1 and 1. If correlation between two variables in close to -1 or 1, it means that those variables have strong negative or positive linear relationship with each other. That is change is one variable is linearly impacted by change in the other variable.

- From the above correlation table, it can be observed that age of the car and its price have a high negative correlation of **-0.732** which indicates that as the age of a car increases, the price of that car decreases.

Correlations

	price	Miles_run	BHP	AGE	gear_auto	gear_semi
price	1	-.623**	.432**	-.732**	.137	.270**
Miles_run	-.623**	1	.010	.590**	.104	-.189
BHP	.432**	.010	1	-.114	-.080	.045
AGE	-.732**	.590**	-.114	1	.067	-.081
gear_auto	.137	.104	-.080	.067	1	-.034
gear_semi	.270**	-.189	.045	-.081	-.034	1

**. Correlation is significant at the 0.01 level (2-tailed).

REGRESSION ANALYSIS

Regression analysis is used to determine which variables are efficient in explaining the price of the car.

We build the regression model using SPSS software. Our dependent variable that is the variable that we aim to predict is the price of the car and variables such as miles run, performance (BHP), age and gear box are the independent variable that are the variables that explains the dependent variable. After running the regression model, we see that all the independent variables are significant in explain the dependent variable that is price.

While building the model, we cannot accommodate categorical variable in the regression model as it is, so we need to convert the variable gear box into a variable called dummy variable which takes values 0 or 1 where 1 indicates presence of that category and 0 indicates absence of that category.

The most parsimonious model is

Price=12141.161 + ((-.065) *Miles run) + (51.624*BHP) + ((-1185.275) *AGE) + (2778.248*gear_auto) + (2841.412*gear_Semi) + e

From the model it can be understood that for each unit increase in miles run, the price of the car decreases by 0.065, for every unit increase in performance (BHP), price would increase by 51.624, for every unit increase in age, price would decrease by 1185.275 and so on.

As the Sig value of all the independent variables are less than 0.05, the model is **parsimonious**

RESIDUAL ANALYSIS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.902[a]	.813	.803	1141.286

a. Predictors: (Constant), gear_semi, gear_auto, BHP, AGE, Miles_run

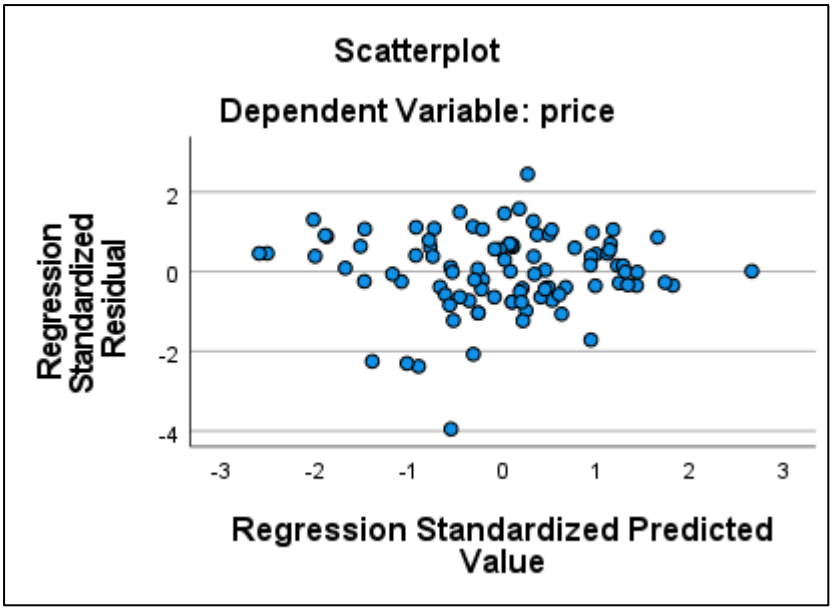
b. Dependent Variable: price

The quality of the model is determined from the coefficient of determination that is R square and adjusted R square. The adjusted R square of the model is 0.803 which means that the **80.3 %** variability in the price is explained by the independent variables. As the adjusted R square is more than 0.7, we can say that this is a good model.

Adequacy of the model

Model adequacy is verified from the following assumptions

1. **Multicollinearity**: from the correlation matrix, we can confirm that there is no multicollinearity between the independent variables.
2. **Independence of residuals**: From the scatter plot between standardized residuals and standardized predicted value, we can see that residuals are randomly scattered which satisfies this assumption.
3. **Homoscedasticity**: Residuals are not showing a specific pattern in the scatter plot which means that this assumption is satisfied



4. **Normality**: From the histogram and normal p-p plot we can see that residuals are normally distributed.