STATISTICAL ANALYSIS III – FINAL REPORT 19333982

INTRODUCTION

The world we live in today is extremely fast-paced. Everyone is always moving from one thing to the next, barely giving themselves time to take a break. This is due to a number of things, one being the speed at which information is capable of travelling nowadays. But another is due to the speed at which humans as a species can now travel from A to B. Historically, people were born and reared in a town, they would grow up there, live there, raise a family there and eventually, probably be laid to rest there, never having left that little town. Nowadays however, it's a completely different story as people travel the world in a matter of days. This revolution in travel has been brought about by developments in all sorts of vehicles, but the most important vehicle to most of us as human beings is the motor car.

Cars are a huge part of our lives as human beings, and as a result of this, most of us, at some stage in our lives, will probably find ourselves in the position that we are buying or selling a car. The purpose of this report is to analyse the Used Cars dataset, and use the information I uncover to aid me in answering the following questions:

- ∇ *Can an appropriate selling price of a car be accurately predicted using existing information on publicly listed cars?*
- ∇ *Can the prices of some cars be more accurately predicted than others?*

This type of study has of course been completed before, and even quite recently for example in the International Burch University in Sarajevo in 2019[1] . These scholars used machine learning methods to try to accurately predict the price of used cars, and they achieved this with an accuracy of 87.38%. I will aim to try and match or improve this level of precision. Of course, selling price of a car can sometimes come down to immeasurable things such as sentimentality or history, however for the purpose of this study I will be excluding these types of factors from my model, and focusing solely on the average market prices of cars. Also, I would like to state that for the majority of this study, I will be operating under the assumption that each type of car that I use to train my regression models is very similar to the others, and that they are all priced in a similar way such that there is no cause for worry with regard to imbalanced data. Afterwards, I will separate my training sets out by type of car, and compare the accuracy of predictions of selling price across the different models of car.

DESCRIPTION OF DATASET

My initial dataset contained just under 100,000 used car listings, comprised of 9 different brands of car each with several different models. This was much too large of a dataset to analyse, so in order to cut down on this size and also increase the model's usefulness, I decided to focus in on particular models of car. I chose three different models for my analysis, the Ford Focus, Ford Fiesta, and Ford Kuga. I chose these three different models as they are obviously all from the same manufacturer, Ford, and they had a high number of observations each. My plan was to analyse the three car models combined into one dataset, and also separately, and hence compare the results of the three models of car with each other. Note that all data was collected from cars listed for sale in 2020.

This new improved dataset contained 10,712 observations, divided into 4,787 Fiesta's, 3,918 Focus's and 2,008 Kuga's. This made the dataset a lot more functional. It also contained 9 variables, with 6 numeric, 2 binary integer and 1 factor. These variables were as follows:

- Price – Numeric – The price at which the car was put up for sale ***(Dependent)***
- Year – Numeric – How many years it has been from when the car was listed to when the car was registered, e.g. registered 2017, year = 3.
- Transmission – Binary Integer – The type of gearbox within the car, either Manual or Automatic
- Mileage – Numeric – The number of miles that the car has driven in its lifetime
- Fuel Type – Binary Integer – The type of fuel used in the car's engine, either Petrol or Diesel
- Tax – Numeric – The amount of road tax due to be paid annually for the car
- Miles per Gallon – Numeric – The number of miles the car can drive per gallon of fuel used
- Engine Size – Numeric – Volume of fuel and air that can be pushed through the car's cylinders, or in everyday terms "the size of the engine".
- Model – Factor with 3 levels - Specifies the model of the car in question, either 1 for Fiesta, 2 for Focus or 3 for Kuga

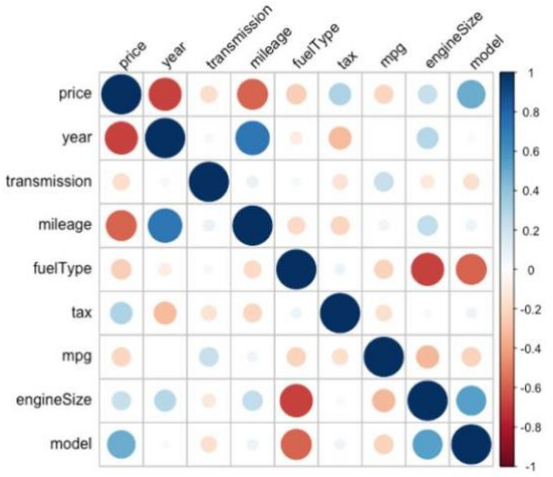
INITIAL CLEANING:

There are a number of alterations that I made to my previous dataset, firstly in an attempt to make the study more accurate, but also to whittle down the numbers from the original 100,000. First, I transformed the year value from the year the car was registered to the age of the car at listing, as described above. This was done to make the model more relevant today, as this data was collected in 2020 and cars depreciate at quite a high rate. Then I removed any cars which had "Electric", "Hybrid" or "Other" fuel types as this made the study a bit more complicated due to a number of reasons, for example the fact that Electric cars sometimes have zero road tax. There were also a minority of cars which were Electric, Hybrid or Other so I concluded that removing them would not have had a major impact on the dataset. Also, petrol and diesel cars are (unfortunately for the environment) the most common seen on our roads today, so I decided it would not majorly decrease the usefulness of this study. Similarly, I decided to remove those cars with "Semi-Automatic" or "Other" transmission as this would make the dataset easier to use.

In order to clean the data, I removed any and all NA values in the dataset, as well as any invalid inputs or unrealistic outliers. For example there were some cars for which engine size, miles per gallon and tax had not been recorded and was input as zero (petrol and diesel cars), and one car which was said to be registered in 2060, and so I decided to remove these observations as well in order to improve accuracy. From my knowledge on the topic, I tried to remove any outliers which I deemed unrealistic or not useful to this study. For example, one car was two years old and priced at £54,995, which is far too high a price for any of these three models of car. I would guess that if it was not a mis-entry, this car was modified or improved somewhat which makes it less relevant to my research question. The purpose of this study was to aid me in buying/selling my own car and so these observation was not wholly relevant.

Finally, I changed the transmission and fuel type variables to numbered values so that I could fit the model to this dataset. For clarity, I gave the "Automatic" cars a transmission value of zero, and the "Manual" cars a transmission value of 1. Similarly, I gave the "Diesel" cars a fuel type value of zero and the "Petrol" cars a fuel type value of 1. Lastly, in the combined dataset with all three models, I assigned the three models of car a different number 1, 2 or 3 for Fiesta, Focus and Kuga respectively.

EXPLORATORY ANALYSIS



For my exploratory analysis, I decided firstly to create boxplots of each of the numeric variables by model to examine their distributions and note any extreme outliers. I also decided to plot each of these variables against the price, in order to try and spot any relationship that may exist. These plots can all be found in the appendix. I then created a Spearman correlogram of all of the numeric variables in the dataset to check for multicollinearity (right). I used Spearman correlation due to the fact that this approach is non-parametric and it assumes that data is measured on a scale that is at least ordinal[2] . Both of these fit my dataset as I do not know what the underlying distribution of my variables was and it did not violate the assumption. I also checked the variance of each variable and covariance of each pair of variables in order to further understand the data.

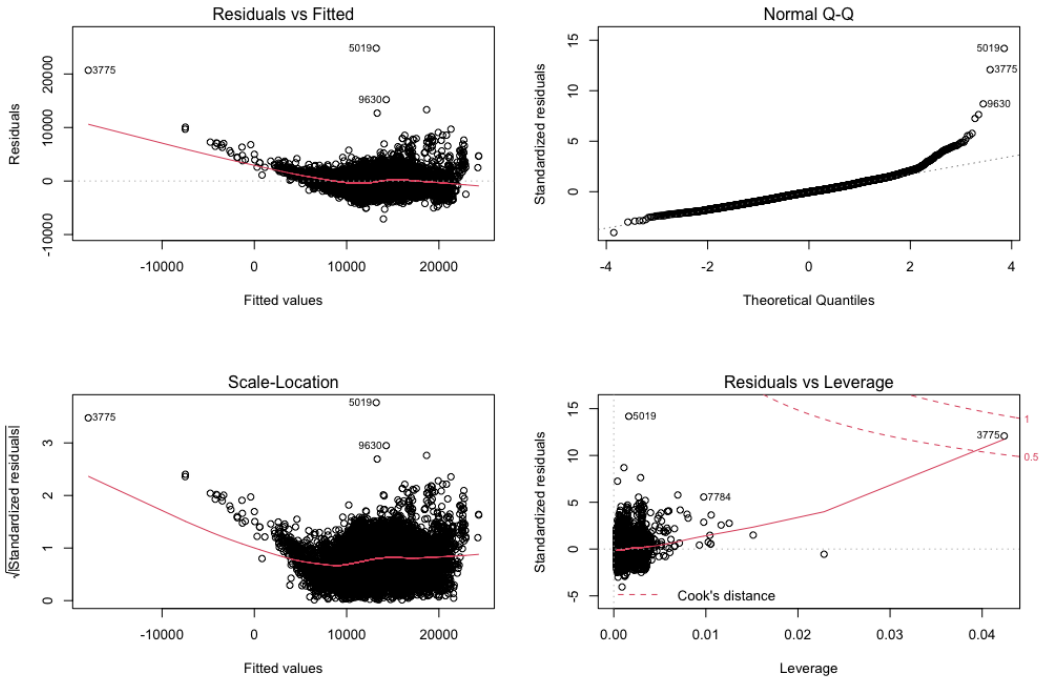
Firstly, upon examination of the boxplots I did not find any of the outliers to be unrealistic and so I made no more exclusions based on this data. After examining the Spearman correlations between each pair of variables, I did notice a high correlation between certain pairs which could indicate multicollinearity, particularly between year & mileage and engine size & fuel type. However, fuel type is a binary integer i.e. it can only take the values 0 and 1, so hence I do not believe that it is likely there is much relevant correlation between these two variables. Multicollinearity can cause insignificant p-values and coefficients[3] so I will keep this potential threat to the model in mind as I progress through this project.

METHODS

In order to fit each linear model, I started by splitting my data into a training and a testing set. The purpose of this was to allow me to assess the goodness of fit of my model using cross-validation with a completely new set of observations, and hence calculate the accuracy of my model with real listed vehicles. I decided to use an 80:20 split for my data, so I took 80% of the Fiesta, Focus and Kuga observations and labelled this the training set, the remaining 20% making up the test set. I then fit three separate models, each one slightly different from the others. Lastly, I split up the data by model and fit three more models, one for each type of car.

MODEL 1:

I used linear regression to describe the relationship between each of the independent variables and the selling price, and analysed the output. The R-squared value was quite high, up at 0.8136, meaning that about 81% of the variation in the price was being described by each of the variables in the dataset. However, due to the fact that we had quite a few independent variables in this study, the Adjusted R-squared is more appropriate as this tells us if there are variables that are not describing any of the variation whereas R-squared does not. In this case, the Adjusted R-squared was slightly lower, at 0.8134, but almost the exact same as the adjusted R-squared. Also, the p-value for this model was extremely small, which indicates statistical significance. I decided to analyse the diagnostics plots of this model, which can be seen below:



The first plot is a scatterplot between residuals and predicted values. There are some unusual points at the top of the graph, and the points are not distributed evenly. The second plot is a normal probability plot. It will give a straight line if the errors are distributed normally, but the points at the top deviate quite a bit from the straight line, especially those numbered at the very top. The third plot indicates the spread of the residuals. Similar to the first, there are some unusual points across the top, and the points are not distributed evenly. The fourth plot helps us to find the inferential points in the dataset. As we move from left to right here, the decreasing spread of residuals indicates heteroskedasticity, and there is one point (labelled 3775) lying outside of Cook's distance. I examined this outlier and discovered that it is a 22 year old Fiesta which means it would now be categorized as a classic vehicle[4]. It also only had a mileage of 37,000, indicating that it had barely been driven at all in its lifetime considering that the average Irish petrol motorist drives 10,500 miles per year[5]. All of this would suggest that this car would most likely not be priced the same way as a standard, newer car. For this reason, I removed this point from the dataset for any further analysis.

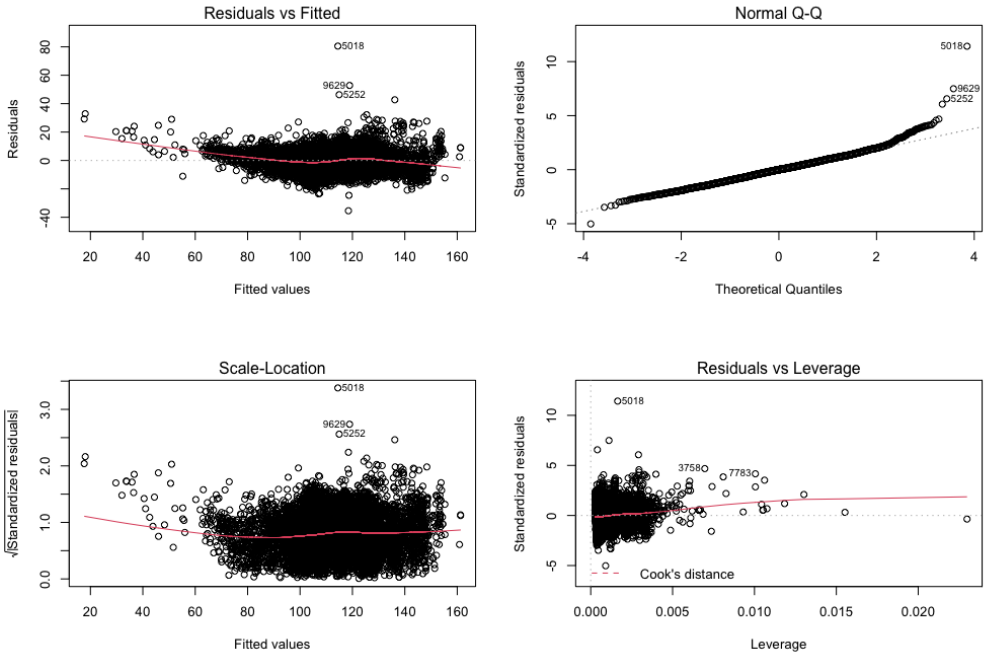
The data point with the largest residual value, numbered in each of the graphs as 5019, is a 3 year old Ford Focus with only 197 mileage. From my knowledge on the topic I can say that there is a good chance this was logged incorrectly, however it is not impossible that this could be correct. Due to the fact that this data was collected in 2020, it is possible the person rarely used their car for about a while, and then the Covid-19 pandemic struck so they were working from home and so the mileage on the car never increased. Hence it is not necessarily an unrealistic observation and for this reason, along with its minimal leverage in the fourth graph, I will leave this data point in the set.

Using cross-validation to assess the accuracy of this model yielded an mean absolute error (MAE) of 1757.011. I calculated this figure by using the model to predict a price for each vehicle in the test set, and then calculated the errors for each observation and found the average. This means that the predictions from this model were accurate within approximately £1,757.01 of the actual price. Considering the average price of a car in the test set is about £11,000, this level of accuracy is fairly good but probably not precise enough to be useful in the real world. Hence, I tried to improve this

model further, applying a transformation to the price by finding the square root, and re-ran the model.

MODEL 2:

The results of this regression were improved slightly from the last model. Our adjusted R-squared value increased to 0.8389, meaning 83.89% of the variance in the log price is described by the variables in question. I then decided to use the *'stepAIC()'* function in R, which uses stepwise regression and aids with variable selection by minimizing the AIC. Unfortunately this did not improve the model any further, however this told me that there were no unnecessary variables being used



and worsening the fit of the model. This second model yielded an MAE of 1241.272, which is a valuable improvement on the first model and makes this a lot more useful.

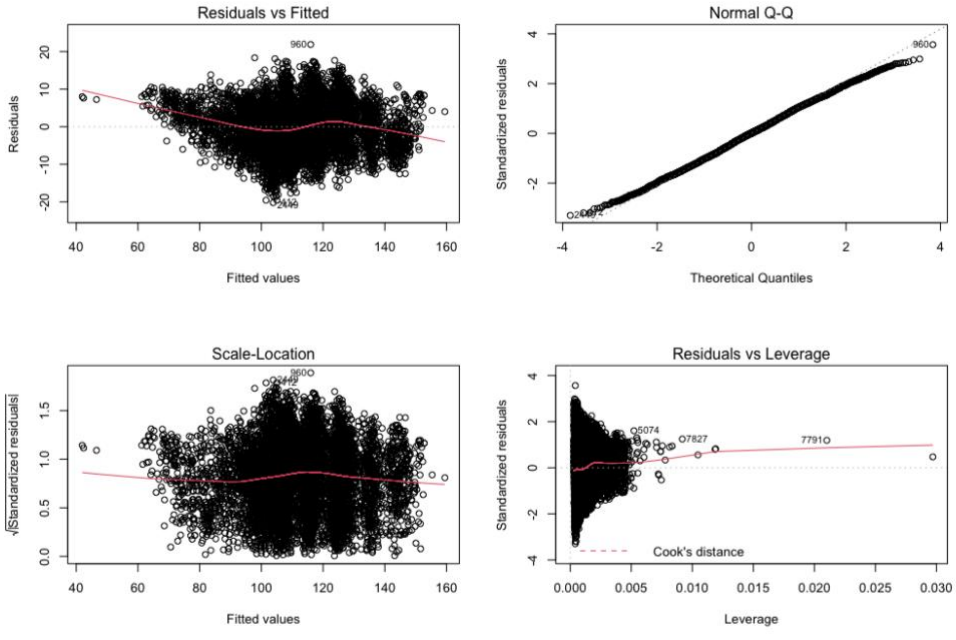
Examining the diagnostic plots for the second model above, we again see some outliers which could be causing inaccuracies. The Normal Q-Q plot still shows sizeable deviation from the straight line with the higher values, numbered 5018, 9629 and 5252. This time there are no points lying outside of Cook's distance which is positive. I will now attempt to eliminate the rest of the skewedness from the Q-Q plot by eliminating more of these outliers using the general rule of thumb that data points with a Cook's Distance of 3 times the mean should be removed from the dataset.

MODEL 3:

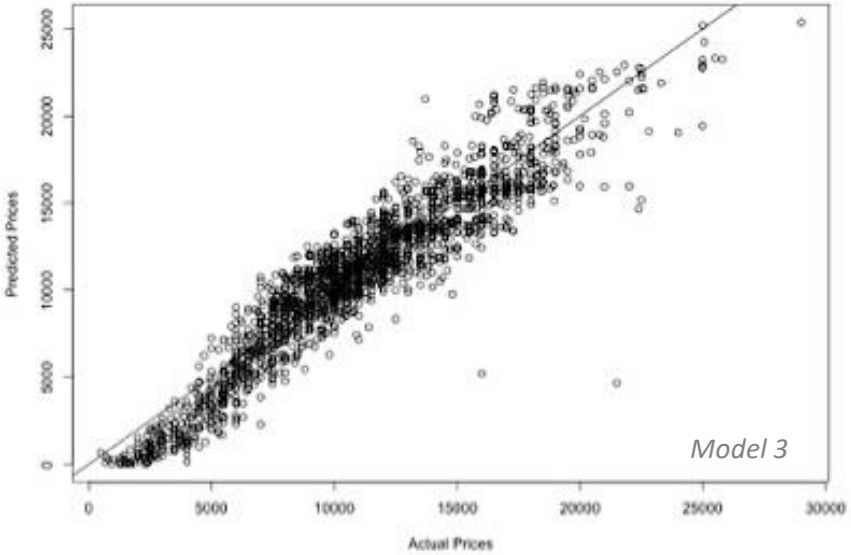
After removing these outliers, the Adjusted R-squared increased to 0.8622, or 86.22%. The output of this model can be seen here. It is evident that the p-value is extremely low and residuals are also approximately centred around 0.

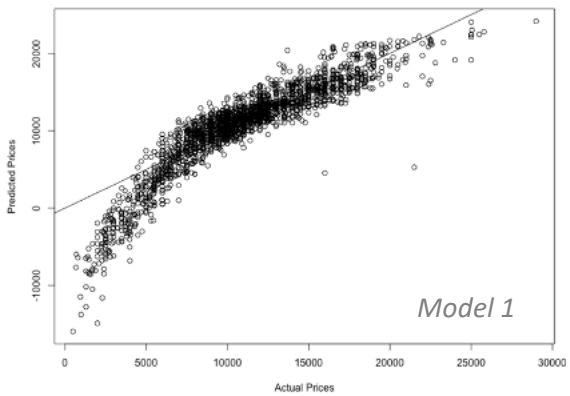
```
Call:
lm(formula = price ~ year + transmission + mileage + fuelType +
    tax + mpg + engineSize + model, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max
-20.216  -4.334   0.060   4.332  21.878

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.019e+02  1.597e+00  63.830  < 2e-16 ***
year         -7.243e+00  7.594e-02 -95.380  < 2e-16 ***
transmission -4.590e+00  3.433e-01 -13.372  < 2e-16 ***
mileage      -2.668e-04  7.129e-06 -37.428  < 2e-16 ***
fuelType      6.580e+00  4.126e-01  15.948  < 2e-16 ***
tax          -2.334e-02  2.080e-03 -11.221  < 2e-16 ***
mpg          -5.130e-02  1.410e-02  -3.637 0.000277 ***
engineSize    1.943e+01  4.173e-01  46.574  < 2e-16 ***
model         8.274e+00  1.462e-01  56.580  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.136 on 8077 degrees of freedom
Multiple R-squared:  0.8623,    Adjusted R-squared:  0.8622
F-statistic:  6322 on 8 and 8077 DF,  p-value: < 2.2e-16
```

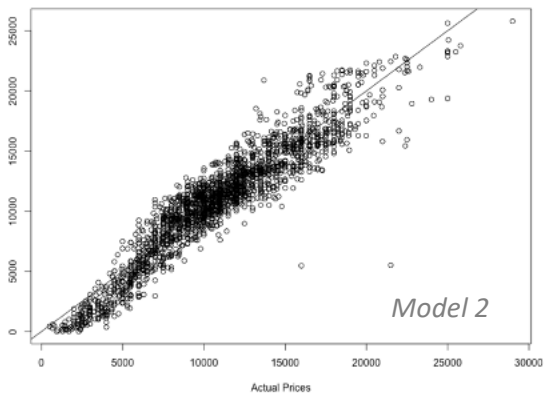
From the diagnostics graphs below we can see that points are spread more randomly than before on the Residual vs Fitted plot and the Scale-Location plot. The Q-Q plot also shows that much more of the residuals follow the straight line, and the points with more extreme deviation are no longer there. Lastly, there are no points outside of Cook's distance in the Residuals vs Leverage plot.



Examining the plot to the left, which shows the Actual values vs the Predicted values for the test set in model 3, is also a very useful and also simple tool for assessing the goodness of fit of a regression model. The perfect model would have all of the points lying exactly on the 45 degree line through the graph, indicating that every predicted value is exactly the same as the actual price. This graph shows how good of a fit our model is, as all of the points are quite close to this line, if not on it, and they form a band centred around this line moving up the graph. The same graphs can be seen below in respect of the first two models, and it is noticeable how inferior the first model is in particular. The graphs of the second and third models are quite similar, from some of the points on the graph especially those closer to the line, and the fact that it has a higher adjusted R-squared value, we can infer that model 3 is slightly more accurate.







Separating into Different Models of Ford:

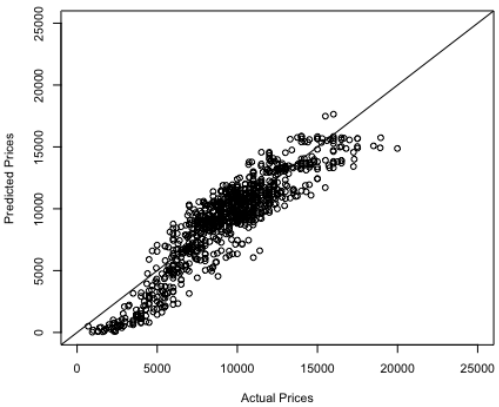
I will now separate out the different models of Ford cars into different datasets. As stated at the very beginning, so far throughout this study I have operated under the assumption that these three models of Ford are all very similar to one another, and that there was no cause for worry with regard to imbalanced data. Again, the purpose of this was to determine whether some models of car can have their prices predicted more easily/ accurately than others.

	P-Value	R^2	Adj R^2	MAE	RMSE	Mean Price
Fiesta	<0.01	0.7325	0.732	1257.53	1547.41	9344.57
Focus	<0.01	0.8979	0.8977	1117.46	1482.15	11091.02
Kuga	<0.01	0.8719	0.8713	1317.74	1717.25	14667.08

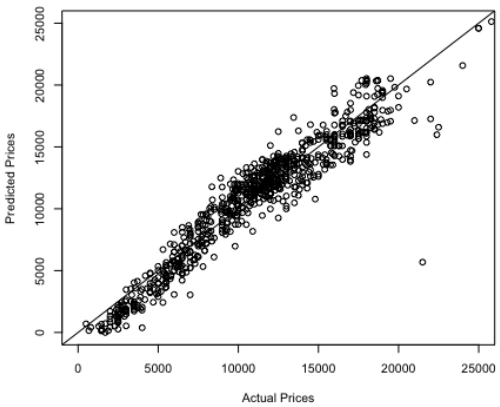
The table on the right shows outputs of the three different linear regression models, for Ford Fiesta, Focus and Kuga respectively. I trained these models using the same data as the third model from earlier, except separated out by model of car. This allowed me to keep any outliers identified previously excluded from my analysis. I then separated out the test set from the earlier model, into the three sections again, and predicted new prices for each brand of car using its respective linear model.

The Adjusted R-Squared Value for Ford Focus's is the highest one we have seen yet, and interestingly the Adjusted R-Squared for Ford Fiesta's is the lowest. This is not what I anticipated at the beginning of this study, as we have more data to train our machine for Ford Fiesta's than any other model of car. Perhaps there is more variance regarding the type of Fiesta itself, for example more options for engine size, than there is for the other brands and this could be the cause of the added difficulty in predicting an accurate selling price for this car. The MAE value for Focus's is also notably low, almost allowing predictions for selling price to an accuracy of ± £1,000.

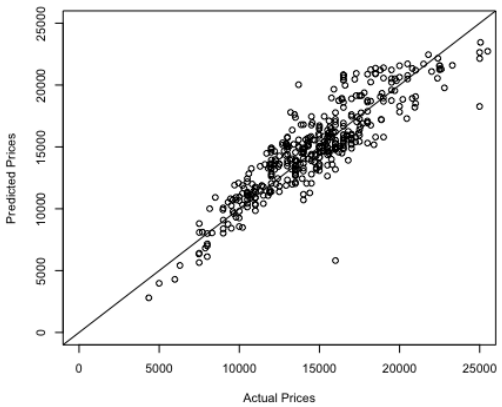
Fiesta



Focus



Kuga



The three graphs above give a side-by-side comparison of the actual prices of each of these test values vs the predicted prices given by each model. Based off both of these comparisons, the table and the graphs, it is clear that the predictions for Ford Fiesta's as a whole are not as accurate as the other two models of car. Focus's and Kuga's both show graphs with smaller residuals, as almost all of the points are extremely close to the line.

CONCLUSIONS

Overall, I believe that I have been quite successful in completing the tasks that I set out with in the beginning. To summarise, I aimed to analyse the Used car dataset which I downloaded from Kaggle, and use the observations within to answer the following two questions:

- ∇ *Can an appropriate selling price of a car be accurately predicted using existing information on publicly listed cars?*
- ∇ *Can the prices of some cars be more accurately predicted than others?*

I started by exploring the data itself, and making any exclusions I saw fit. When I had finished cleaning and preparing my data, I then fit three different linear models each with its own different attributes. Using these three models I predicted the selling price of a variety of different cars with a fairly high degree of accuracy. I then separated out the three different models of Ford (Fiesta, Focus, Kuga), and sought to explore of this would help to achieve more accurate predictions. For the most part, it did. The selling price of Ford Focus's and Ford Kuga's could be achieved with an impressively high degree of accuracy using models trained exclusively with data from those types of car. Ford Fiesta's on the other hand could not. In fact, the least accurate model that I fitted, very interestingly, was trained using data from only one type of car.

I found this study extremely interesting to carry out. The fact that the Fiesta model was so much less accurate than the other models was something that fascinated me, and so I decided to briefly research this more, as I am certainly not a car expert. I quickly found that there are 6 different subclasses of Ford Fiesta, each starting at different price ranges, compared to only 5 different versions of the Ford Focus and 4 different versions of the Ford Kuga. Perhaps this range of options and the differences between them is the reason that prices of Ford Focus's are harder to predict. I think that perhaps with more data, specifically which type of Fiesta each observation is, this could be uncovered.

As stated in the introduction, I was aiming to match or beat an accuracy of 87.38%, as achieved by academics in the International Burch University in Sarajevo[1] . Although I am not 100% certain of which metric they are using to compute this accuracy, I did manage to achieve an adjusted R^2 value of 89.77%, so I will call that a success. In fairness, their model is a little more advanced as it takes into account the brand, the model and a few more variables than mine, however they had more advanced regression techniques at their disposal.

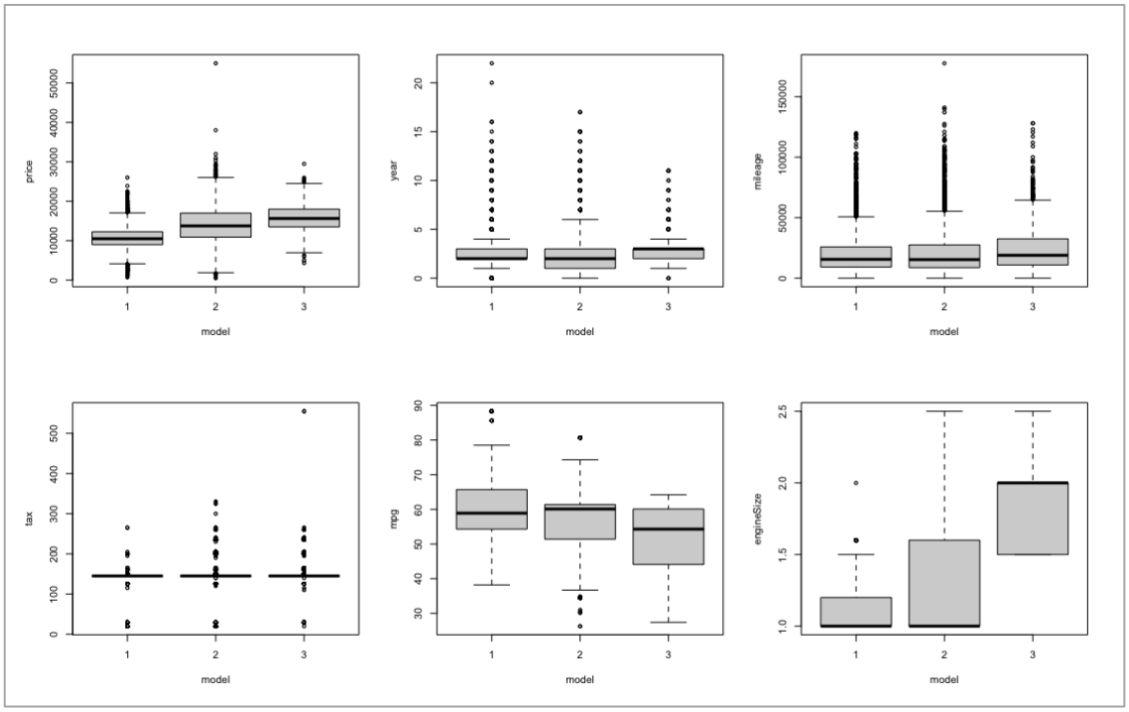
In answer to my two-part research question above, an appropriate selling price of a car can be accurately predicted using existing information on publicly listed cars. Also, based on the results of my study, the prices of some models of car can be predicted more accurately than others, and this probably depends on the different options there are for that specific model. I have no doubt that with the use of more advanced regression methods and a more in depth analysis, more information could be uncovered and price predictions could be made with an even higher degree of accuracy.

REFERENCES

1. https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf
2. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/kendalls-tau-and-spearmans-rank-correlation-coefficient/
3. https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a
4. https://www.campion.com/blog/car-insurance/when-is-a-car-classic/
5. https://www.ccpc.ie/consumers/cars/car-clocking/

APPENDIX

Boxplots by model



Variables vs Price