Statistical Analysis III – Project

Part Two: Exploratory Analysis

Introduction: The purpose of this exploratory analysis is to uncover characteristics of my dataset which will help me to better understand the data, and guide the analysis that I plan to carry out to answer my research question, which is: *Can an appropriate selling price of a car be accurately predicted using information on publicly listed cars? Can prices of some cars be more accurately predicted than others?*

Description: My working dataset can be found in the attached csv file "UsedCars.csv". The initial dataset contained just under 100,000 used car listings which was far too large of a dataset. 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 is to analyse the entire dataset, and also the three car models 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 contains 10,713 observations, divided into 4,787 Fiesta's, 3,918 Focus's and 2,008 Kuga's. This makes the dataset a lot more functional. It also contains 9 variables, with 6 numeric, 2 binary integer and 1 factor. These variables are as follows:

- Price – Numeric – The price at which the car was put up for sale ***(Independent)***
- 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

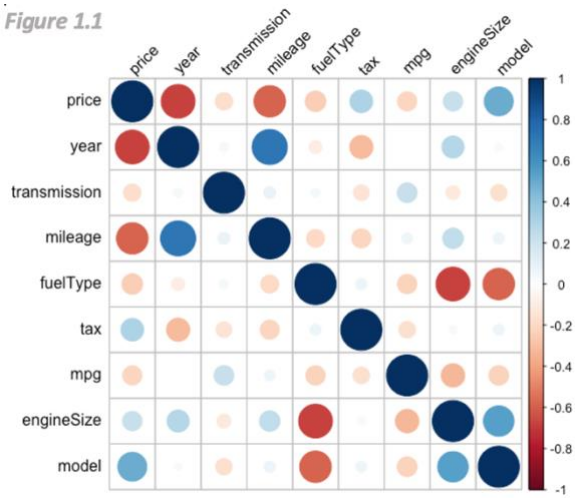
Exclusions & Cleaning: There are a number of changes that I made to my previous dataset, firstly in an attempt to make the study more accurate, but also to whittle down the numbers from the original 100,000.
First, I added the brand column and combined the different data frames using the rbind() function in R.
Next, 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 2017 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) 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.
Thirdly, I decided to remove the model column as I am analysing each of these models of car separately so it bears no relevance.

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 (even 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. And 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.

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 then created a Spearman correlogram of all of the numeric variables in the dataset to check for multicollinearity (Figure 1.1). 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 (1). Both of these fit my dataset as I do not know what the underlying distribution of my variables are and it does 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.

Figure 1.1



Results/Conclusion: 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 (2) so I will keep this potential threat to the model in mind as I progress through this project.

References

1. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/kendalls-tau-and-spearmans-rank-correlation-coefficient/
2. https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a