Statistical Analysis III – Project

Part One: Dataset & Background

My chosen dataset can be found at the following link:
https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

This dataset contains details gathered about 100,000 used car listings, separated by car make. Each file is named according to the make of car it describes.

There are also two files included which the creator called "uncleaned" files. These contain more raw data which has not been properly tidied up and edited for the purpose of statistical analysis. The creator has simply included these files to show others the process he used to "clean" the data, and hence is not very relevant to my project. I will therefore be excluding these files from my analysis.

As well as this, the Ford Focus and Mercedes C-Class files are subsets of the Ford and Mercedes files respectively. The creator made these files initially, which they outline on Kaggle, and then extended them to form the other newer files. For this reason I will also be excluding them from my analysis so as not to count these vehicles twice.

This leaves me with 9 files, representing 9 different types of car:
Audi, BMW, Ford, Hyundai, Mercedes, Skoda, Toyota, Vauxhall and Volkswagen.

And each file contains 9 columns, representing 9 variables relating to each vehicle:
Model, Year, Price, Transmission, Mileage, Fuel Type, Tax, Miles per Gallon, Engine Size

As stated above, the files themselves have already been "cleaned" by the creator, therefore there are no other exclusions to be made for the purpose of this analysis. For example there are no rows containing NULL values or that are missing certain metrics.

There are a number of vehicles which have a tax value of 0, which I initially thought may indicate that certain vehicles did not have the tax amount recorded. However, upon further investigation I spotted that we can still accept this as an accurate observation due to the fact that some cars require zero tax to be paid. This can be for a variety of reasons, usually related to the CO2 emissions of the vehicle.

In terms of my research question, I wish to predict a suitable selling price for a vehicle of any of the above brands in relation to the current market, as well as investigate if there is an optimal time to sell a given car, for example a specific age where resale value drops significantly, or a specific mileage.

At first glance, this dataset seemed as though it could require the use of clustering, however the fact that the cars are already separated (or clustered) into their relevant car brands means that clustering would not be a very interesting approach in my opinion. Instead, I have decided that I will use linear regression to identify the relationship between selling price (my chosen dependent variable) and the other independent variables as listed above. Furthermore, I hope to be able to use my model to predict a relevant selling price for any of the examined brands of car. This will be a very useful tool in my opinion, at least for me personally, as I am hoping to sell my own little car when I have finished college and (all things going well) upgrade to a newer/ more improved vehicle. This project would not only allow me to determine a suitable selling price for my current car in relation to the market, but also decide whether a given new car is undervalued or overvalued.

Lastly, I will attempt to use my findings to compare the selling prices of different types of cars, and the difference in the effects of each independent variable on the selling price of each car brand.

I chose this dataset as it contains an abundance of information to explore, and also because I think it will provide interest and useful findings. There is of course potential that the market for these cars will have changed since this data was gathered about one year ago, however I think that this potential issue could be overcome by using relative pricing, for example by viewing the year of the car not as a quantity in itself but rather as “number of years old”. In this case, cars registered in 2020 would be zero years old, 2019 would be one year old, and so on.

If one wished to be even more thorough, they could include approximate inflation rates in their calculation of suitable prices for future years. Otherwise the data could be simply updated and the same model used. As this project continues, I will examine some of these possibilities in greater detail.

To conclude, I am very happy with my chosen dataset and look forward to exploring it further in the coming weeks.