Principles Of Data Science:   
An Analysis of Vehicle Pricing in Relation to Given Features

Data Understanding and Preparation

*1.1: Meaning and type of Features*

**Quantitative Features:** Public Reference, Mileage, Reg Code, Year of Registration and Price.

**Qualitative Features:** Standard Colour, Standard Make, Standard Model, Vehicle Condition, Body Type, Crossover Car and Van and Fuel Type.

**Mileage:**

The distance the car in the listing has been driven, measured in Miles. In relation to other features, if the vehicle has an older Year of Registration, it is likely its mileage will be higher. Also, certain car models will typically have higher mileage, such as cars used as commercial vehicles. Any Missing values in this dataset will be set to the mean Mileage of other vehicles from that Year of Registration.

**Vehicle Condition:**

This feature shows if the car is new or has previous owners. Related to the Year of Registration, older cars are more likely to have had a previous owner.

**Year Of Registration:**

This feature is when the vehicle was first registered, measured as a Year. In the case of new cars, the Year of Registration is left blank, and so will need to be filled in with the year the dataset is taken from. Anomalous values can be repaired within most cases using the vehicle's Registration Code, which is also given in the dataset.

**Price**

The asking price for purchase of the vehicle, measured in GBP. Price is related to all other features in some way, as these features are what decide the desirability of the vehicle, and in turn, the price that the seller can reasonably ask to receive in return.

*1.2: Analysis of Distributions*

Figure 1: Year of Registration Boxplot

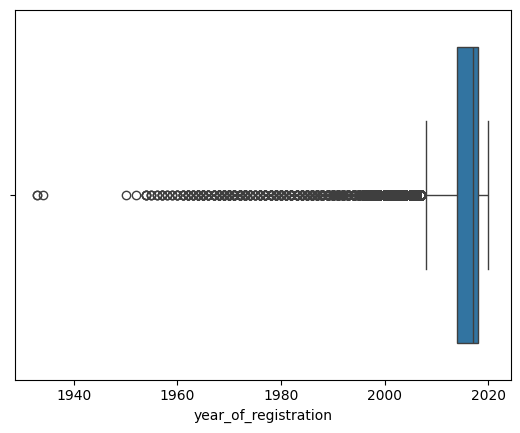
This Boxplot clearly demonstrates the outliers of this feature, which need to be either repaired or removed during Data Cleaning.

Figure 2: Year of Registration Histogram

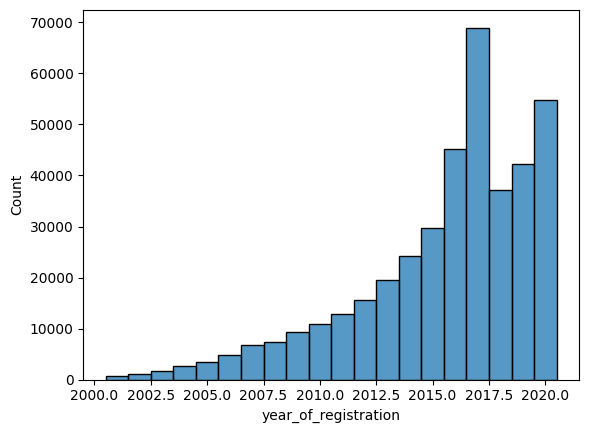
This Histogram has had the lowest X-Axis value limited to 2000, to accurately show the distribution of values. This data follows a somewhat exponential distribution.

Figure 3: Price Histogram

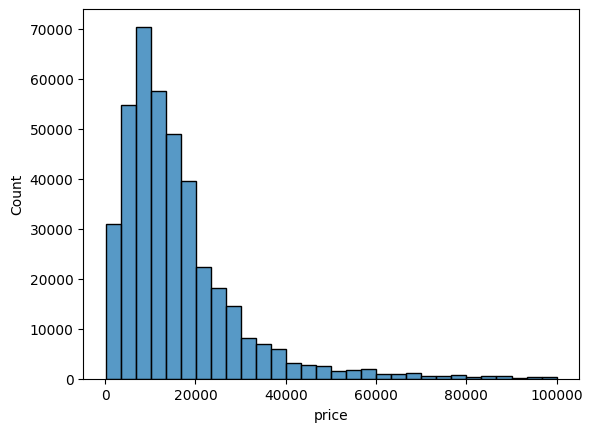
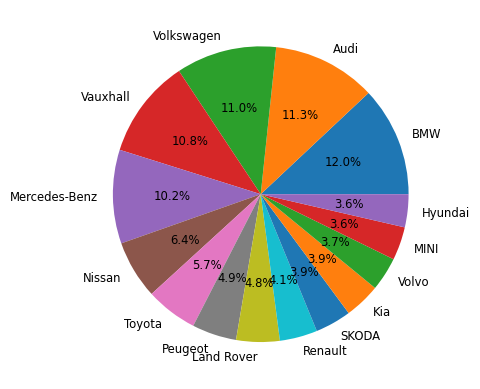
The Price histogram clearly demonstrates the small IQR of the price feature, and its similarity to a normal distribution around its mean value.

Figure 4: Standard Make Pie Chart

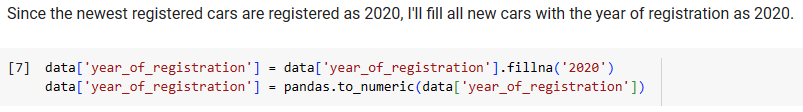
A pie chart of the 15 most listed Makes, with the Top 5 makes holding a market share of almost 50%.

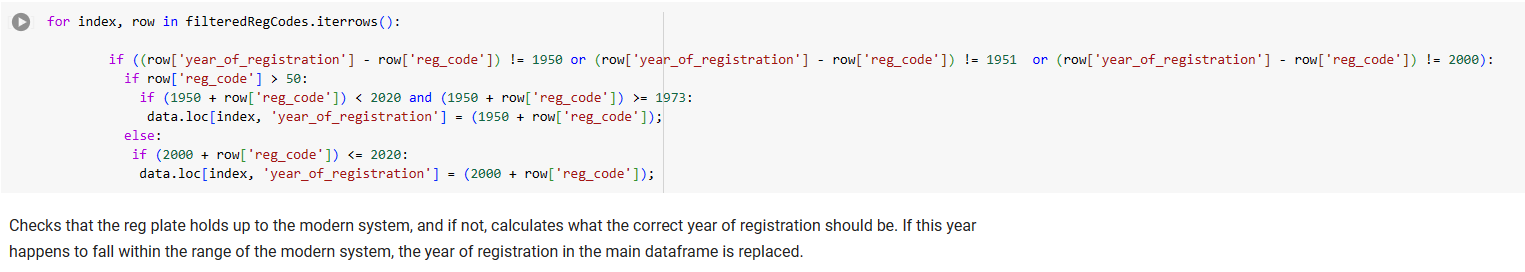
Data Pre-Processing

*2.1 Data Cleaning*

Year of Registration

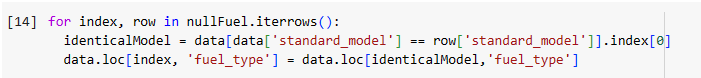
When viewing a sample of the data, I noticed a blank value for the Year of Registration of a new car and realized that this will pose an issue when it comes to analysis of the Year of Registration's effect on the price. This was fixed by going through the dataset and filling in all blank values with the year this dataset was taken from (2020).

Afterwards, when viewing a description of the Year of Registration feature, I noticed that the Minimum value was 999. This was an anomalous value, and using nsmallest(), I viewed the other lowest values and realized that incorrect years of registration must be a common error. This was fixed by checking that the reg plate holds up to the modern system, and if not, calculates what the correct year of registration should be. If this year happens to fall within the range of the modern system, the year of registration in the main dataframe is replaced. This method is not perfect, as non-integer reg codes and those using the old style from 1973 are not affected, but these are only a small amount of those in the dataset.



Fuel Type

This Column also contains multiple Null results, although I found a quicker and simpler way to correct this data. The following algorithm, when given a null value, finds a car of identical model and fills in the missing fuel type.

This algorithm successfully filled in 93% of null fuel types, more accurately than a random assignment would be. This resulted in reducing the total number of null fuel types from 601 to just 43, all in cases where an identical model with a valid fuel type could not be located.

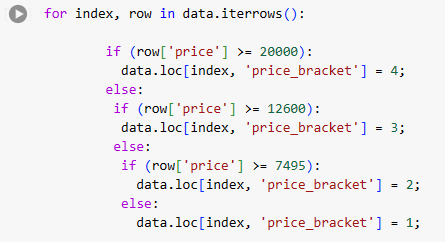
Body Type

This column had a similar issue to the Fuel Type column, and therefore I could use a slightly adapted version of the previous algorithm to also repair the data found here. The null values were reduced from 837 to 76, a success rate of 11%.

*2.2 Feature Engineering*

Price Brackets

The First thing I will be adding to the database is a 'Price Brackets' column. This will allow easy use of data exclusively within a particular quartile of Price, important when looking at the concentration of categorical values at different price points.



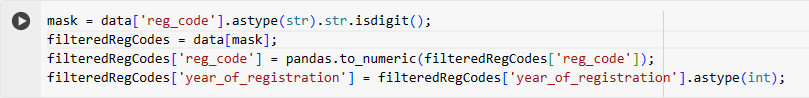
Crossover Car and Van Removal

As Crossover Car and Van is applicable for so few listings in the dataset, I will remove it from the data in order to save time when processing data.



*2.3 Subsetting*

Registration Codes Mask



During the removal of the anomalous values from the Year of Registration column, I needed to remove all string values from the Registration Code column, and so, used a mask to ensure the only remaining values were digits. This facilitated the algorithm seen earlier to validate the Registration Year of the vehicles in the listings.

Null Value Selection

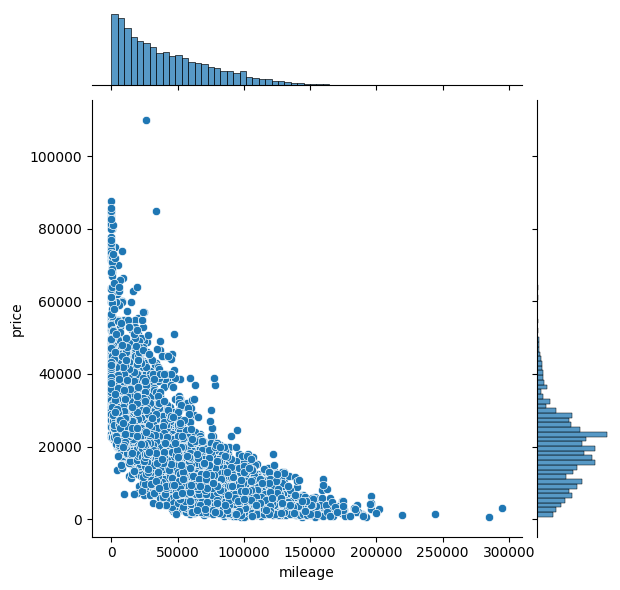
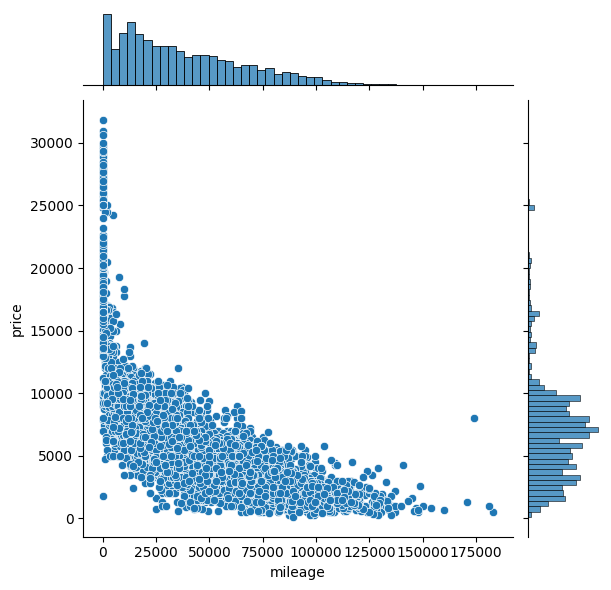
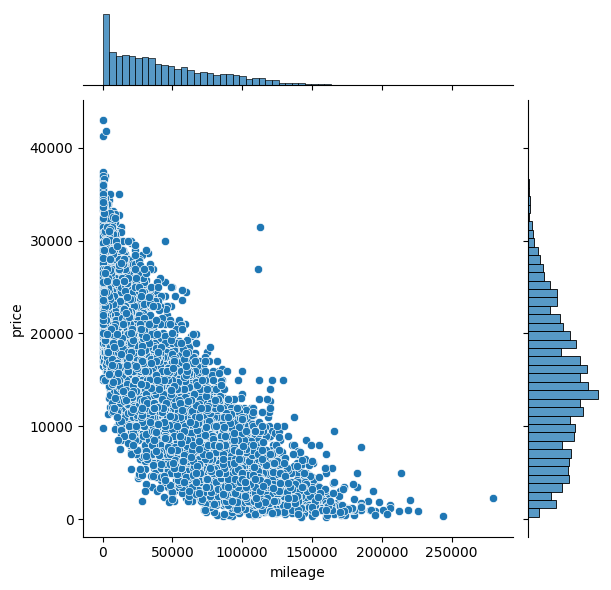


Also during the Data Cleaning process, it was necessary to create a subset of exclusively listings with null values in the Fuel Data and Body Type columns. This allowed the previously seen algorithm to replace these null values in those columns.

Analysis of Associations and Group Differences

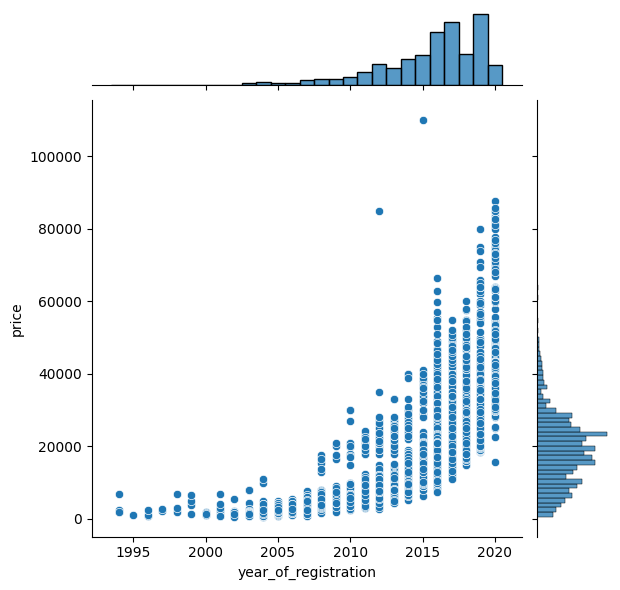
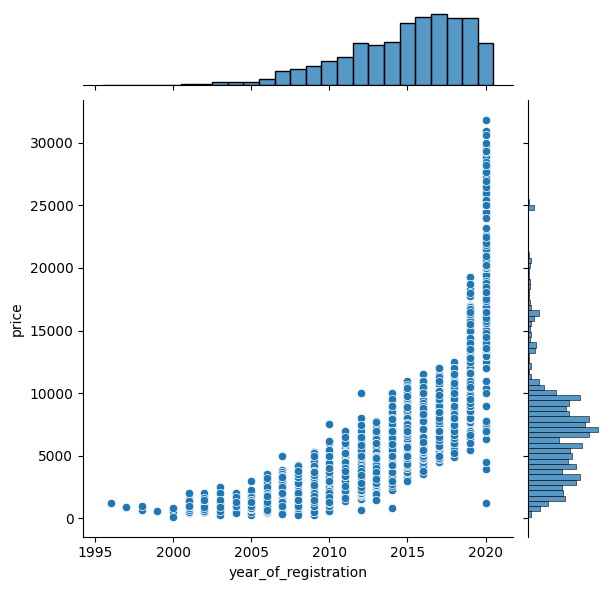
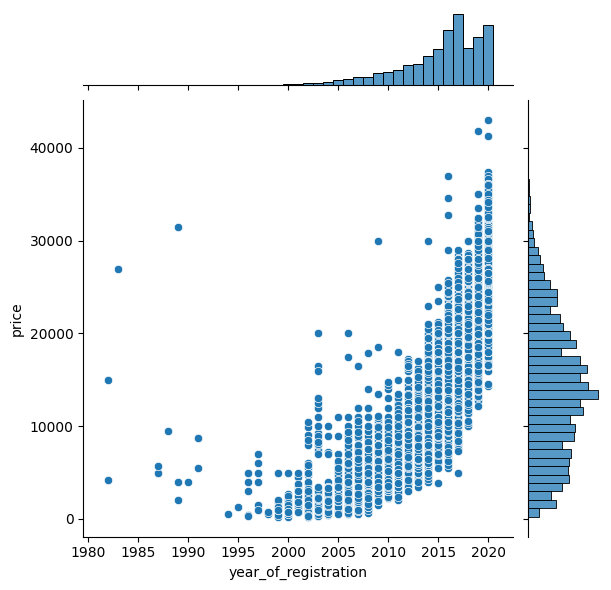
*3.1 Quantitative-Quantitative*

Price-Mileage Jointplots



3 Jointplots, all showing the relationship between Price and Mileage within a single model of car. I used the 3 most popular models, from left to right being a Volkswagen Golf, Vauxhall Corsa, and Mercedes C-Class. These graphs show that Mileage has a strong negative correlation with Price within a particular model. This implies that a lower Mileage is an important contributor to increasing the desirability of a given vehicle.

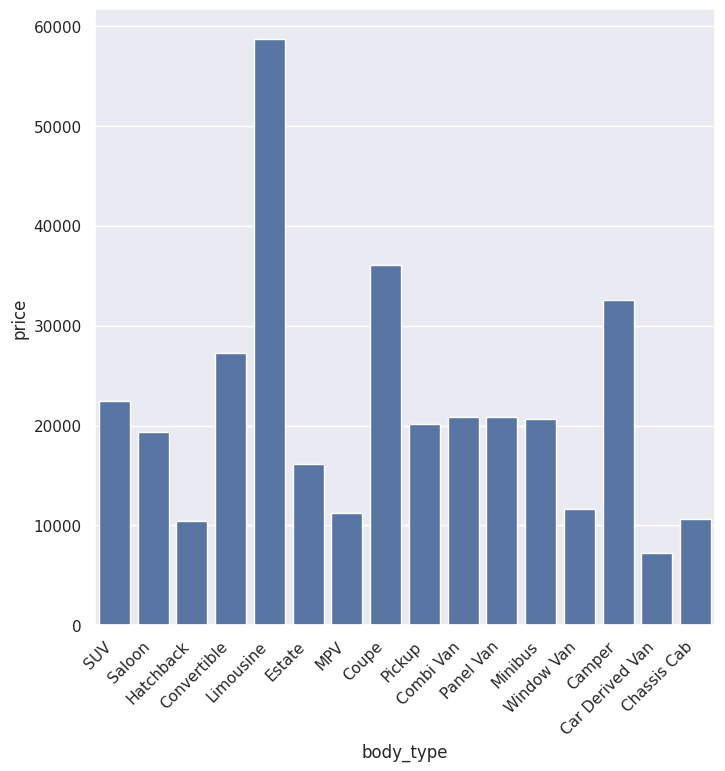
Price-Year of Registration Jointplots



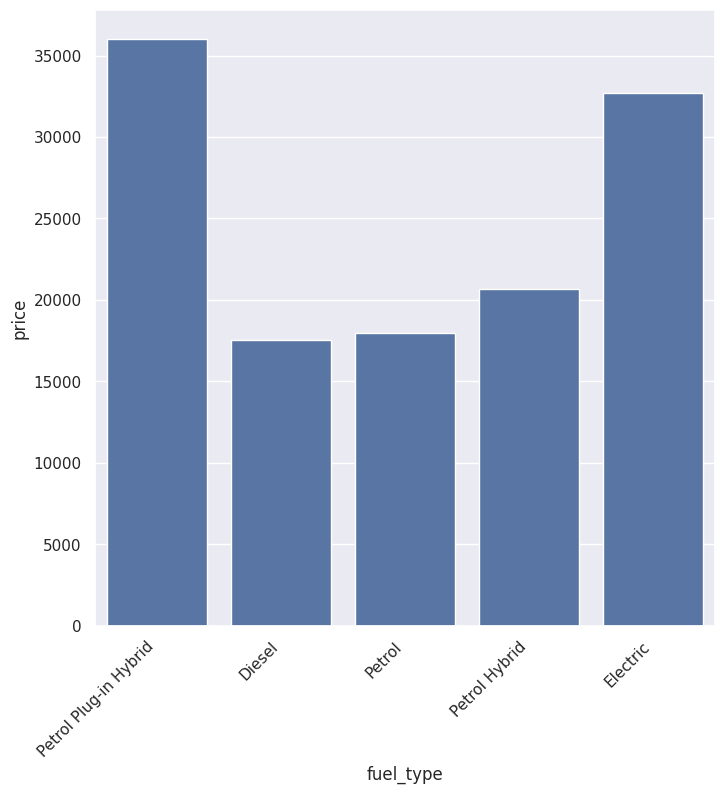
3 more Jointplots, all showing the relationship between Price and Year of Registration within a single model of car. I used the same 3 most popular models, as a lack of correlation here would imply that mileage is a more important factor to the price of a vehicle. Instead, what these graphs imply is that whether a car is new or used is a more important factor.

*3.2 Quantitative-Categorical*

Price-Body Type Barplot

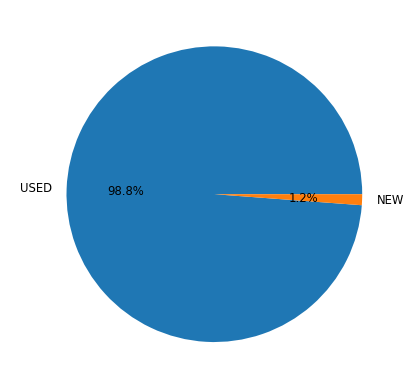
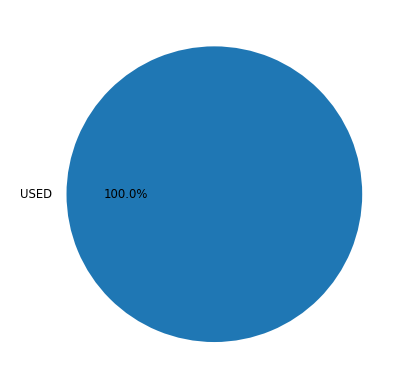
A Barplot displaying the relationship between price and Body Type. This graph shows that Body Type significantly impacts the price of a vehicle, with some Body Types, such as Limousines, being much more expensive. This is likely due to two factors. The desirability of a certain Body Type, and the original cost of production. The high average price of coupes can be somewhat attributed to the presence of high-end sports cars in the dataset, which are mostly coupes.

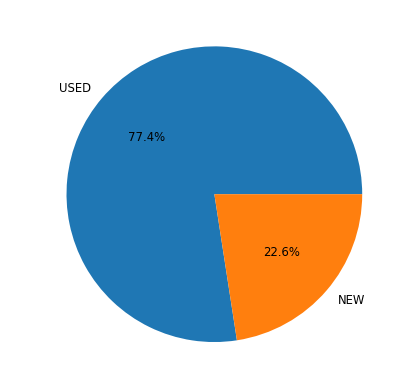
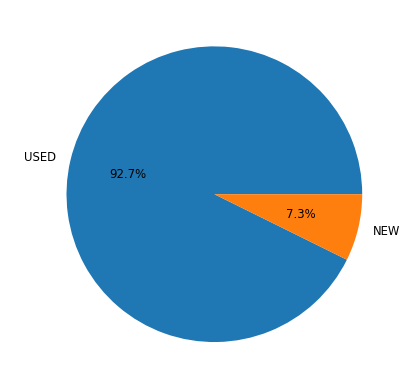
Price-Fuel Type Barplots

This graph shows that Fuel Type has a significant effect on price, with Electric and Petrol Plug-in Hybrid vehicles (seemingly) being more desirable than other fuel types. This could also be due to the production costs of these vehicles being higher. It should also be noted that this data was taken from the 5 highest market-share fuel sources, and only data from after all had presence in the market was used (2010, The first Petrol Plug-in Hybrid is registered in the dataset), to avoid any misleading results.

*3.3 Categorical-Categorical*

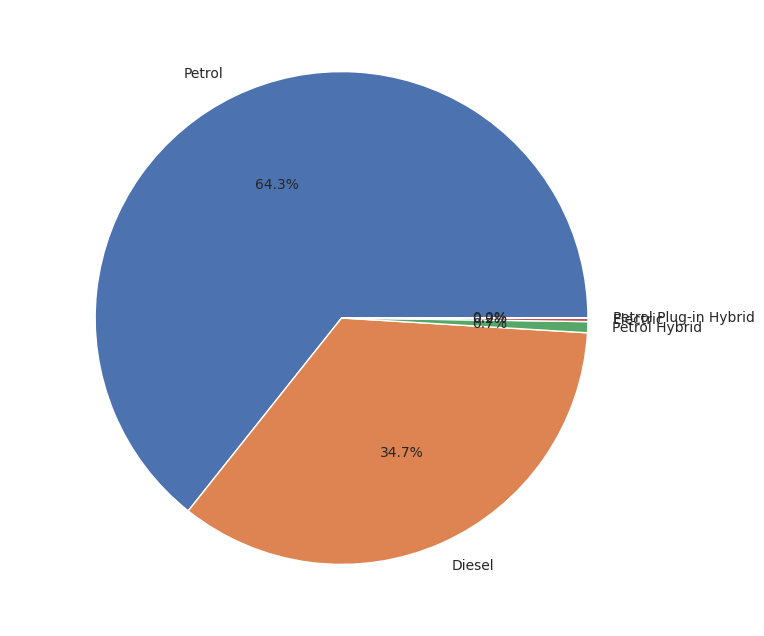
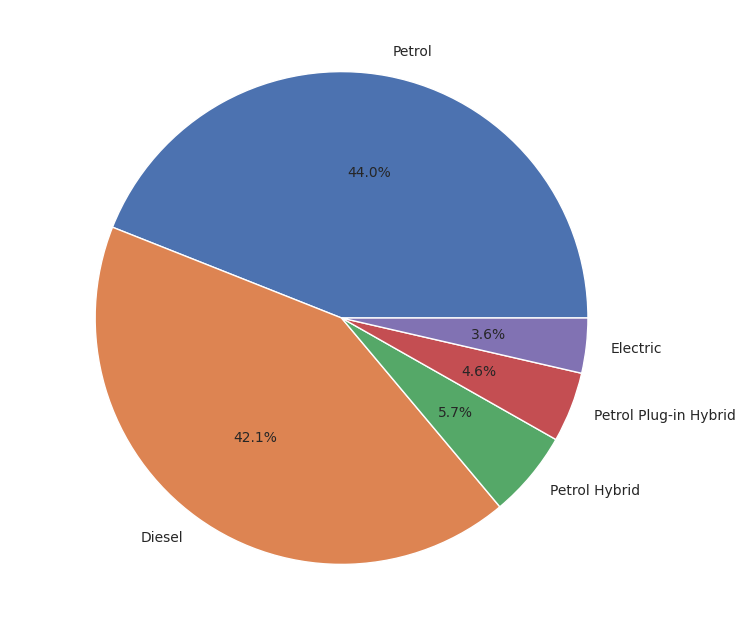
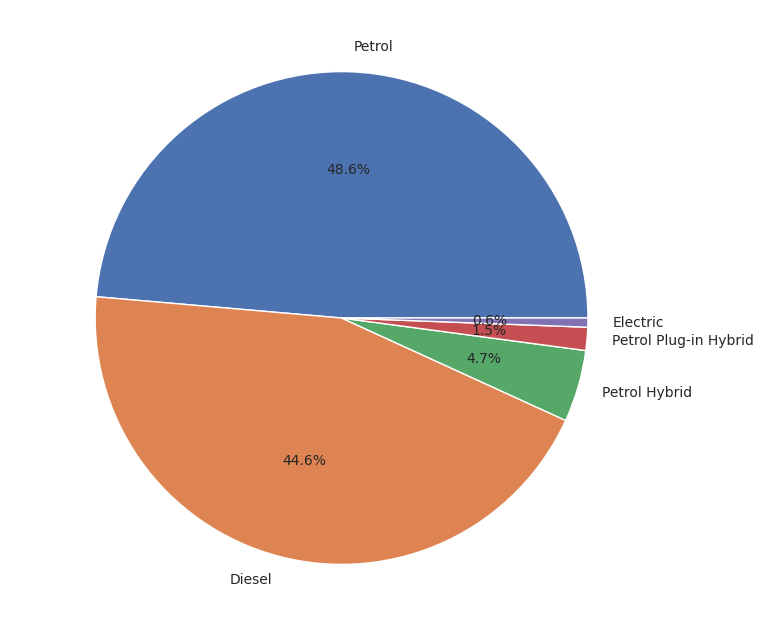
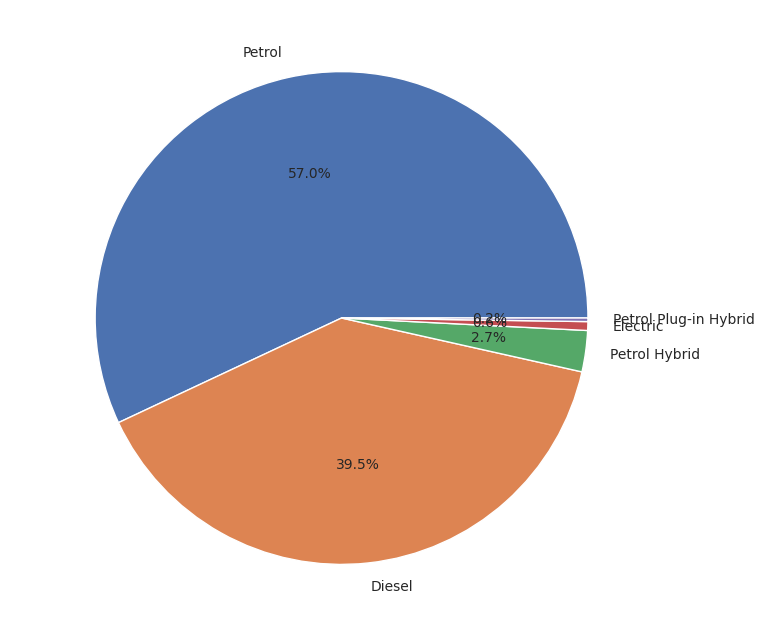
Price Band-Vehicle Condition Pie Charts





These Pie Charts show the portion of New and Used cars available for purchase in each Price Bracket. The Graphs show that Vehicle Condition is a good predictor of price. I find it particularly interesting that new vehicles are completely absent from the lowest price bracket, and also having an extremely low presence in the 2nd price bracket, meaning that a car being new almost guarantees it will be in the top 50% in terms of price. The large presence then, in the highest price category, leads me to believe that Vehicle Condition is in fact the feature **most** indicative of price included in the dataset.

Price Band-Fuel Type Pie Charts



I thought that an interesting grouping to explore would be the market share of different fuel types at different price points. The above Pie Charts cover 2010-onwards to not skew results in favor of petrol and diesel, since the First Petrol Plug-in Hybrid and electric cars in the dataset were registered in 2010 and 2009 respectively. The Pie charts show, from top left to bottom right, the market presence of the top 5 fuel types in each quartile of price.

These Pie charts can show the Fuel Types in different quartiles of Price as we are using Price-Range as a qualitative feature. These charts effectively demonstrate the high cost of electric cars in comparison to the average, showing that most hybrid cars are within the top 50% of cars in terms of price, while most electric cars being grouped in the top 25%.

Not only is this data grouping interesting considering the current government push towards electric vehicles for the average consumer, but the recent EU Climate Policies require all vehicles to run on e-fuels by 2035, meaning that a much more consistent effort to bring electric cars into all domains would be expected. The data clearly shows that owning an electric vehicle is simply not economic for the average consumer, as of the date of collection for this data.