Muhammad Taimur Aftab - 24835406

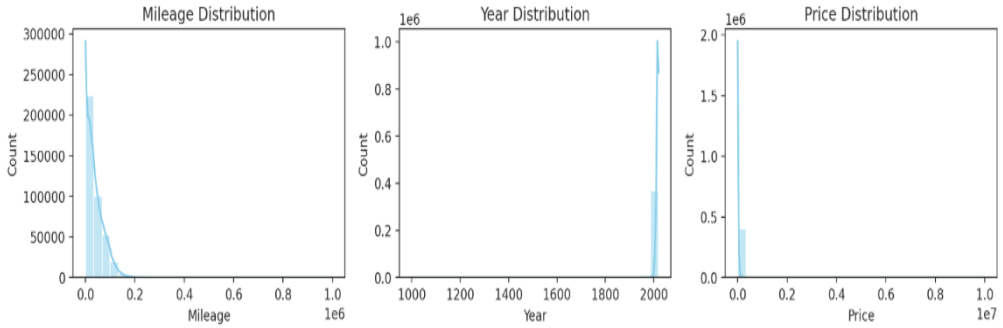
1. Data/Domain Understanding and Exploration

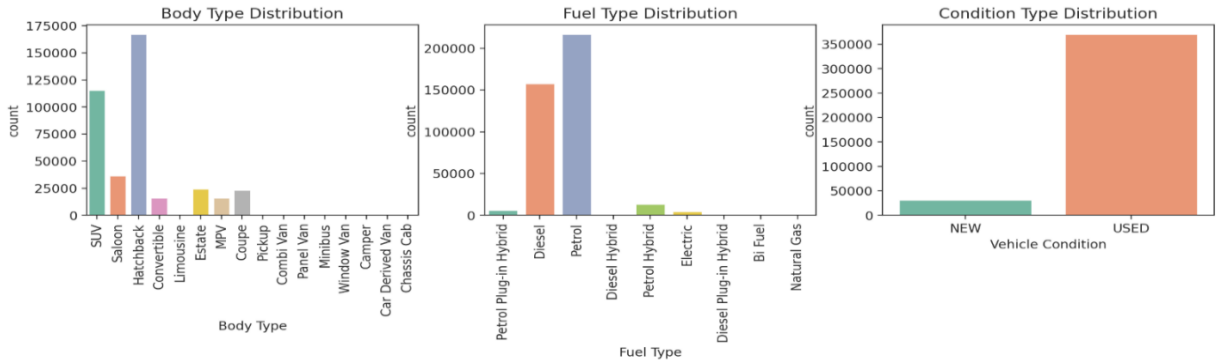
1.1 Meaning and Type of Features; Analysis of Distributions

The data has 402005 rows and 12 features which provides the description of the vehicle in numerical and categorical form. The numerical attributes include mileage, year_of_registration, and price, while the other attributes include standard_make, vehicle_condition and fuel_type, which gives information context of the vehicles. For most of the columns, there are no missing values, though some important features – mileage and year_of_registration, for example – contain a significant number of missing values: 127 and 33,311, respectively. These missing entries could be of concern during modelling most especially for features such as year_of_registration which when included in the prediction of age and price of vehicles will be impacted. The table below has been created to explain the meaning of features and data type in order to help understand the different facets of features.

	Data Type	Description
public_reference	float64	It shows a unique identifier for the car
mileage	float64	It tells the total distance a car has travelled.
reg_code	object	This is the registration code of the car.
standard_colour	object	The color of the car.
standard_make	object	The brand of the car like Audi, BMW.
standard_model	object	The specific model of the car like A5 of Audi.
vehicle_condition	object	The new or used condition of the car.
year_of_registration	float64	The registration year of the car.
price	int64	The price of the car.
body_type	object	Vehicle body type like sedan, SUV, hatchback.
crossover_car_and_van	bool	It displays if the vehicle is a crossover.
fuel_type	object	It shows fuel type that car uses like petrol.

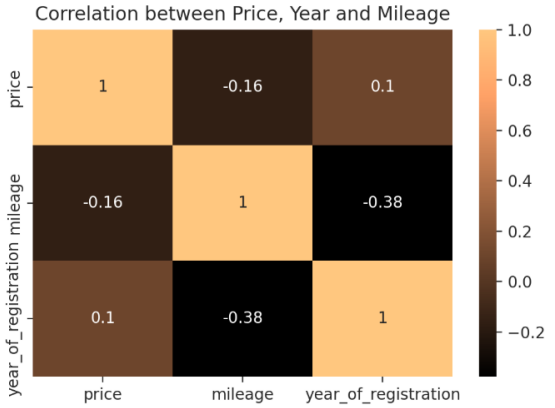
The histograms of features such as Mileage, Year of Registration and Price are presented to illustrate why preprocessing is needed and identify the basic properties of these columns. Mileage is positively skewed meaning most of the cars have low mileage of less than 100000, this indicates that the dataset has low mileage cars. But if we look at the distribution in the long tail up to 1,000,000 miles, we can see that there are outliers that have to be addressed. Most of the years of registration are within 2000 and 2020, probably newer vehicles perhaps there are few cars registered in 1000 and could be due to data entry errors. Similarly, the price distribution is positively skewed as seen by the fact that most vehicles cost less than £20,000, however, there are a few that are priced up to £10,000,000, which probably reflects on luxury vehicles, or just some data entry error. These patterns correspond with trends found in practise, highlighting newer, mid-price, and cars with low mileage. Data transformations such as log scaling or outlier treatment are required to deal with skewness and extreme values which are critical when preparing data for modelling.



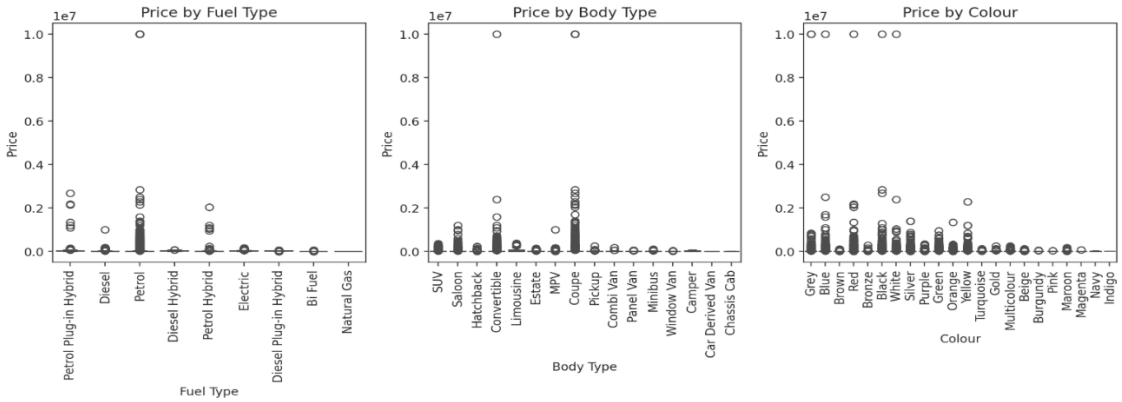


The body type feature indicates that the most common type is the Hatchback whereas the second most common type is SUV, other types such as Saloon and Estate have moderate frequencies while that of the Convertible has low frequency. The fuel type feature is mainly comprised of Petrol and Diesel with very limited representation of Electric, Hybrid, Natural Gas car manufacturers. The vehicle condition feature also displays the dataset comprise most of the used vehicles. These distributions indicate that this dataset are covers preferred body types, traditional fuels, and most used cars.

1.2. Analysis of Predictive Power of Features



When using the correlation heatmap, one is able to notice high correlation between Price, the Mileage and the Year of Registration. Weak negative coefficient of determination between Price and Mileage (-0.16) mean that high mileage is associated with low prices, though a weak coefficient of determination indicate that high value of it can predict price on its own. This is supported by the rather low, yet positive, correlation coefficient of 0.10 between the price and the year of registration, showing that the newer the vehicle the higher its price. Similarly, the moderate negative correlation between the variables Mileage and Year of Registration (-0.38) imply that vehicles with low mileage are newer, thus, these two may not be different from each other and may provide similar information for predictions.



The box plots provide some insight into the relationship of fuel type, body type, and colour with price to establish their effect. Fuel type shows a high likelihood of predicting price because of a direct relationship with prices, luxury and environmentally friendly electric and petrol plug-in hybrids have higher median prices than mainstream petrol and diesel.

Similarly, the body type has a very high predictive value, for example, Convertible and Coupe are related to luxury or high-performance car brands and cost more than, for example, SUVs and Hatchbacks that have a vast variety in price due to the existence of numerous segments.

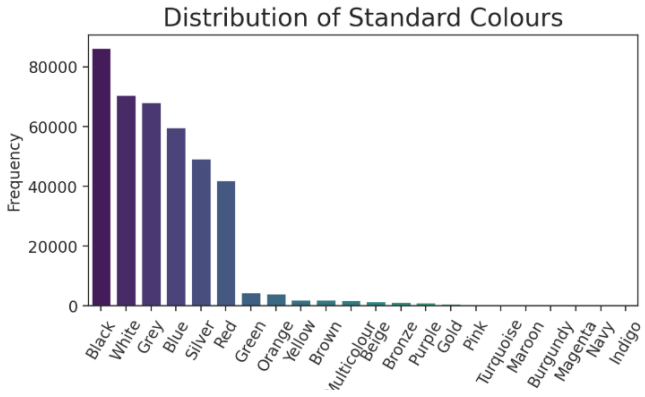
However, colour has a poor capability of predicting, since the distribution of prices by colour such as Black, White and Silver are very close meaning that colour preference does not have any strong influence on the price of vehicles. These relationships show that fuel type and body type are significantly related with market segmentation and consumer value and as such are good indicators of the price. Colour on the other hand has low variability in this context.

1.3 Data Processing for Data Exploration and Visualisation

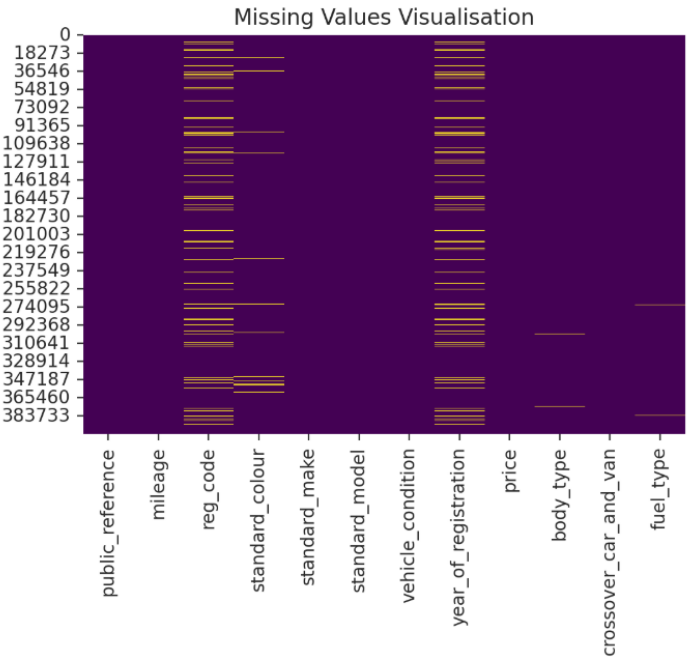
The data set has 402, 005 entries and the numerical features include mileage, year_of_registration, and price. The price ranges from 120 to 9,999,999 with the mean of 12,600. The prices have high variance because the upper limit is significantly higher than the 75 percent. It ranges from 0 to 999,999 and the median value of the mileage is 28629.5, this means that vehicles can be new and old, frequently used. Year of Registration is an ordinal level variable which can take on a minimum value of 999, which is probably an error or possibly a missing value, and a maximum value of 2020.

	public_reference	mileage	year_of_registration	price
count	4.020050e+05	401878.000000	368694.000000	4.020050e+05
mean	2.020069e+14	37743.595656	2015.006206	1.734197e+04
std	1.691466e+10	34831.724018	7.962667	4.643746e+04
min	2.013070e+14	0.000000	999.000000	1.200000e+02
25%	2.020090e+14	10481.000000	2013.000000	7.495000e+03
50%	2.020090e+14	28629.500000	2016.000000	1.260000e+04
75%	2.020100e+14	56875.750000	2018.000000	2.000000e+04
max	2.020110e+14	999999.000000	2020.000000	9.999999e+06

The bar plot shows that among all the vehicles the most frequently selected colour preference is black then white and grey. They are the colours which dominate the selected dataset, which shows that consumers remain conservative and prefer traditional colours. However, Green, Orange and Brown are not often used, even Purple and Pink are rarely used colours. This distribution therefore suggests that the neutral colours may rather have a much larger effect on the resale value and demand of cars than the special ones.



Heatmap displays the missing values where some of the features include like reg_code, standard_colour, and year_of_registration. Year of registration has more missing data which might be as a result of data quality issues, therefore likely to require imputation. Mileage and fuel_type which are the other features have few missing values and do not cause a major impact to the analysis. For two significant features, namely price and vehicle_condition, all values were given and, therefore, they are appropriate for modelling. This analysis shows what needs to be emphasised when it comes to preprocessing of data.



2. Data Processing for Machine Learning

2.1. Dealing with Missing Values, Outliers, and Noise

	0
public_reference	0
mileage	127
reg_code	31857
standard_colour	5378
standard_make	0
standard_model	0
vehicle_condition	0
year_of_registration	33311
price	0
body_type	837
crossover_car_and_van	0
fuel_type	601

In the dataset, there are few features which have missing values namely the year_of_registration (33,311) and mileage (127). In categorical features, reg_code contains 31,857, standard_colour contains 5378, body_type contains 837, and fuel_type contains 601 missing values. This makes it necessary to fill these missing values so as to maintain the credibility of the dataset.

With regards to the missing values for the mileage feature, the mean is selected as the imputation strategy. This approach provides a stable and valid mean of central tendency that ensures the imputed values are correct in representing the feature distribution. After analysing the dataset, there was no new vehicle with null mileage according to the code output. It was possible to make this observation to make the mean only applicable on the used vehicles so that the data set remains wholly uncompromised.

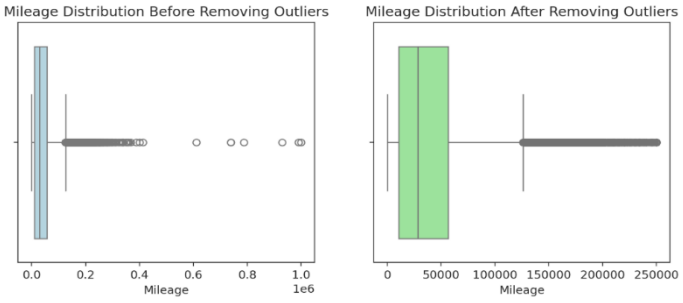
```
[55]  1 any_new_car_missing_mileage = at[at['mileage'].isnull() & at['vehicle_condition'] == 'NEW']
      2 print(any_new_car_missing_mileage)
```

```
Empty DataFrame
```

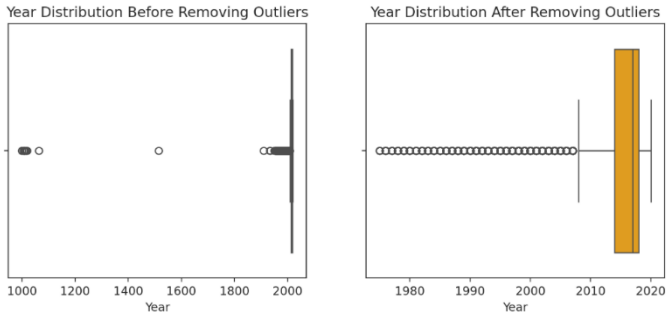
```
vehicle_condition
NEW      31249
USED      2062
Name: count, dtype: int64
```

T There are 33,311 missing values in the year_of_registration feature and new vehicles have 31,249 missing while used vehicles have 2,062 missing values. For the newly registered vehicles their mileage 0 to 100 placed them in the newly registered vehicles hence the maximum year in the data was input manually as the latest registration year. For the used vehicles, the reg_code column was applied for the imputation by mapping it to the year of registration using the UK vehicle registration system. Finally, all the 378 missing values were imputed with the mode year that is 2017 to complete the imputation process.

In the case of fuel_type, body_type, and standard_colour, missing value analysis was conducted and the missing values were randomly imputed based on the frequencies.



As mentioned in Section 1.1, two of the attributes, mileage and year_of_registration have outliers in the dataset. In order to support the analysis of such points, a boxplot was used and it is easy to compare values before and after the removal of outliers. Such outliers were removed to reduce the range to 0 to 250000 miles and the distribution was now realistic and less distorted.



The same observation is seen when the range is restricted to 1975 to 2020, the box plot also gives a reasonable representation of the years of registration of vehicles. These changes improved the quality of both features to ensure that they were accurate for analysis and modelling as well as to be useful in practise.

2.2. Feature Engineering, Data Transformations, Feature Selection

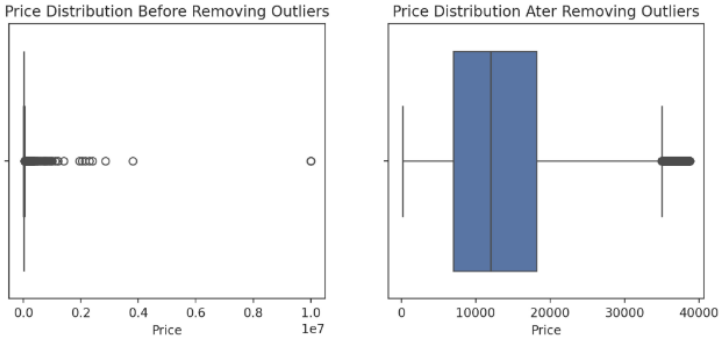
Creating a perfect model requires the removal of unimportant and overlapping columns. Some of these features were omitted to improve the dataset preparation, public_reference was not useful in terms of prediction. reg_code was useful in imputing for year_of_registration but was removed because it was redundant. crossover_car_and_van was removed because it had low variance and weak correlation. standard_model was removed because of high cardinality and difficult to encode and compute.

	mileage	standard_colour	standard_make	vehicle_condition	year_of_registration	price	body_type	fuel_type
0	0.0	Grey	Volvo	NEW	2020.0	73970	SUV	Petrol Plug-in Hybrid

An age_of_vehicle feature was also generated by excluding the max year from the dataset and then subtracting the year_of_registration. After this, the year_of_registration column was removed to avoid the duplication of same feature.

	mileage	standard_colour	standard_make	vehicle_condition	price	body_type	fuel_type	age_of_vehicle
0	0.0	Grey	Volvo	NEW	73970	SUV	Petrol Plug-in Hybrid	0.0
1	108230.0	Blue	Jaguar	USED	7000	Saloon	Diesel	9.0
2	7800.0	Grey	SKODA	USED	14000	SUV	Petrol	3.0
3	45000.0	Brown	Vauxhall	USED	7995	Hatchback	Diesel	4.0

To make the analysis more manageable, the data set was reduced by randomly sampling to half of the original size, down to 200826 rows with 8 columns from the original 402005 rows, 12 columns. This sampling approach allowed the study to be computationally more efficient while still preserving the heterogeneity, as well as the organisational structure of the data.



Following sub setting, the IQR method was applied for the outliers elimination in the price feature. This method retained values that were between 1.5 IQR and rejected others that were not within this range. Lowering the extreme prices was useful in selecting the right measures that ease the distribution and the removal of noise. The resulting dataset was more appropriate for accurate modelling.

Label encoding was not applied because it creates artificial ordinal relationships among nominal variables, One-Hot Encoding (OHE) was applied instead. Target encoding was not performed to prevent the leakage of information. Subsequent to the encoding step, the features of the dataset were isolated from the target variable, and the latter was labelled as X and y. Because the scales of the features are not equal, the features were standardised using StandardScaler so that no feature dominates the others.

Following that, Recursive Feature Elimination (RFE) was applied to the set of features using the linear regression model to identify features that were most relevant to the price prediction. This was especially advantageous in the sense that it served to filter out as many features from the data as possible that were not as influential when making the prediction.

```
Selected Features:
 Index(['mileage', 'age_of_vehicle', 'fuel_type_Diesel', 'fuel_type_Petrol',
       'fuel_type_Petrol Hybrid', 'body_type_Convertible', 'body_type_Coupe',
       'body_type_Estate', 'body_type_Hatchback', 'body_type_MPV',
       'body_type_SUV', 'body_type_Saloon', 'vehicle_condition_USED',
       'standard_make_Aston Martin', 'standard_make_Audi', 'standard_make_BMW',
       'standard_make_Bentley', 'standard_make_Dacia', 'standard_make_Jaguar',
       'standard_make_Land Rover', 'standard_make_Lexus',
       'standard_make_Mercedes-Benz', 'standard_make_Porsche',
       'standard_make_Volkswagen', 'standard_make_Volvo'],
      dtype='object')
```

3. Model Building

3.1. Algorithm Selection, Model Instantiation and Configuration

To predict vehicle prices, three regression models were chosen, Linear Regression, Decision Tree Regression and k-Nearest Neighbour Regression (kNN). Each of the models was selected considering their characteristics and their behaviour when encountering the data samples..

Linear regression is used in an attempt to find out the relationships between the features and the target price as linear. Decision tree is chosen because it can detect nonlinear relationship and interaction effects of the predictors. This is well applicable when handling number data, categorical data and does not need the data to be linear, as well kNN was selected because it functions by looking for close data points to the data point it is tackling.

The kNN Regressor was initialised with its default settings of n_neighbors=5. The number of neighbours to consider when making prediction and consequently decides on how good the model is in capturing local patterns in a given data set.

```
knn = KNeighborsRegressor(n_neighbors=5)
print(knn)
```

The decision tree regressor was also created with random_state=42 to ensure that results are determinable and can be replicated. The random_state parameter is a parameter that was used for splitting of the tree and is used to generate random numbers each time the programme is run so as to produce the same tree.

```
decision_tree = DecisionTreeRegressor(random_state=42)
print(decision_tree)
```

```
DecisionTreeRegressor(random_state=42)
```

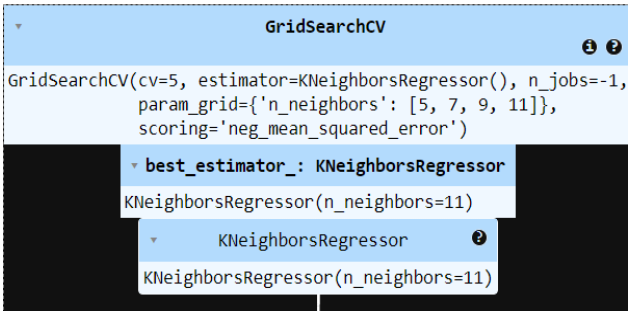
Linear Regression was instantiated in its standard form because no other parameters need to be added into the model. This model assumes for linearity between the features and the target variable and provides an interpretable and simple baseline. The output ensures that the model is ready to identify the linear features in the data and will be compared with other models.

```
linear_reg = LinearRegression()
print(linear_reg)
```

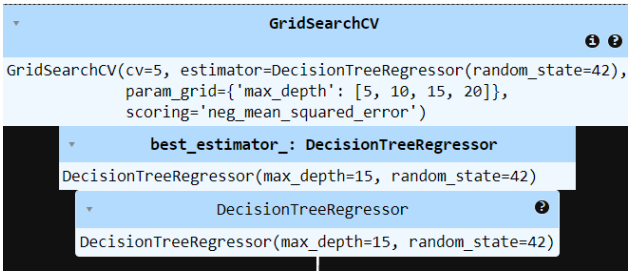
```
LinearRegression()
```

3.2. Grid Search, and Model Ranking and Selection

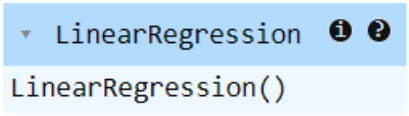
Grid Search was used to identify the optimal hyperparameters for each model by exploring a predefined set of parameter values.



Hyperparameters optimisation of the kNN model was done by using a grid search with 5-fold cross-validation of the hyperparameter n_neighbors with the options of [5, 7, 9, 11]. The chosen configuration was n_neighbors = 11 due to its balance between complexity and generalisation capabilities. Since the finalised model incorporates the best hyperparameter, the test set was used for making the prediction. The tuned kNN model performed well, and the recommended value of n_neighbors = 11 gave good pattern recognition ability but also good ability to generalise to unseen data.



The hyperparameter was also tuned for the decision tree model using grid search with 5-fold cross-validation for the max_depth parameter with values [5, 10, 15, 20]. The max_depth which gave the best performance was found to be 15. After determining the optimised hyperparameter, the model was fitted and used to predict on the test set. It was then demonstrated that the tuned model had reasonable predictive accuracy and max_depth = 15 was determined to be the best parameter that provided capability to capture non-linearity in the data without over-fitting.



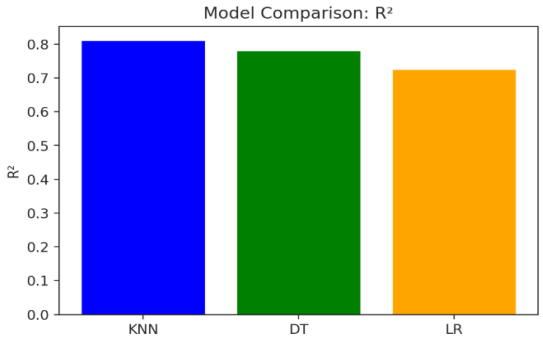
Linear regression was chosen for baseline on purpose in order to see how well it performs since linear regression models the relationship between input features and the target variable as being directly proportional. It is faster to compute, and does not require any form of optimisation of the hyperparameters. However, it is rigid in the sense of modelling because it can only estimate linear relationships and hence is used only as a benchmark of comparison to more complex models.

From the kNN model with n_neighbors = 11, we obtained MSE of 12896982.38, RMSE of 3581.23 and R^2 score of 0.8111. The max_depth tuned as 15 of decision tree model gave MSE as 14806067.40, RMSE of 3,847.86 and R^2 of 0.7832. Linear regression gave an MSE of 18666028.05, RMSE of 4320.41 and R^2 score of 0.7266.

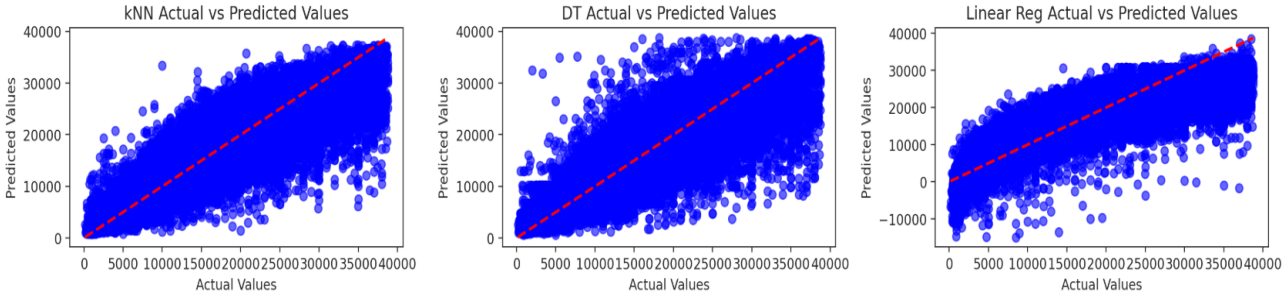
Therefore, the recommendation in this case is the kNN model because of high accuracy, as well as the best performance in terms of patterns' differentiation to new data.

4. Model Evaluation and Analysis

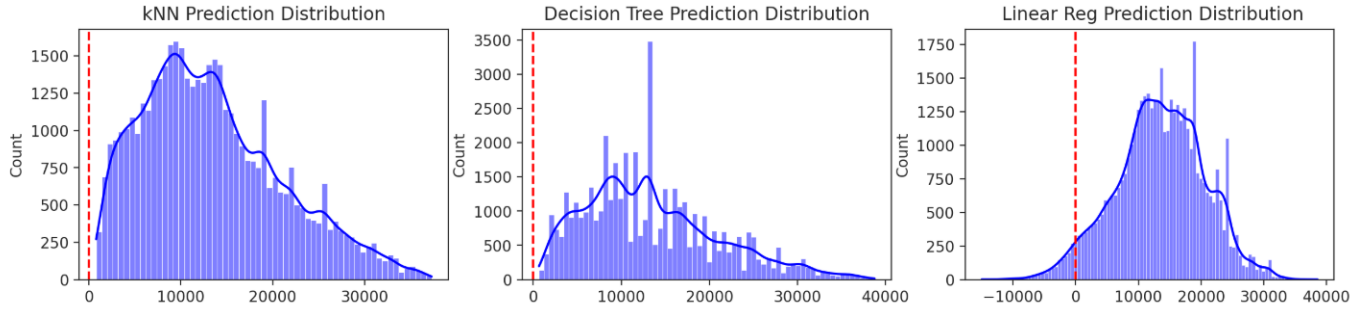
4.1 Coarse-Grained Evaluation/Analysis



The coarse-grained evaluation enabled the comparison of the R^2 of the models to check their ability to explain the variation in the target variable. The kNN model has the highest R^2 of 0.81, which is high enough to confirm good generalisation and good capability of capturing patterns with the best n_neighbors at 11. The decision tree model has R^2 score of 0.78 which shows satisfactory model fit to handle the non-linearity and the model generalisation was slightly compromised because of overfitting even though the model depth was set at 15. Linear Regression has the lowest R^2 score of 0.73 because it tried to establish a linear correlation between the features of the dataset and could not.

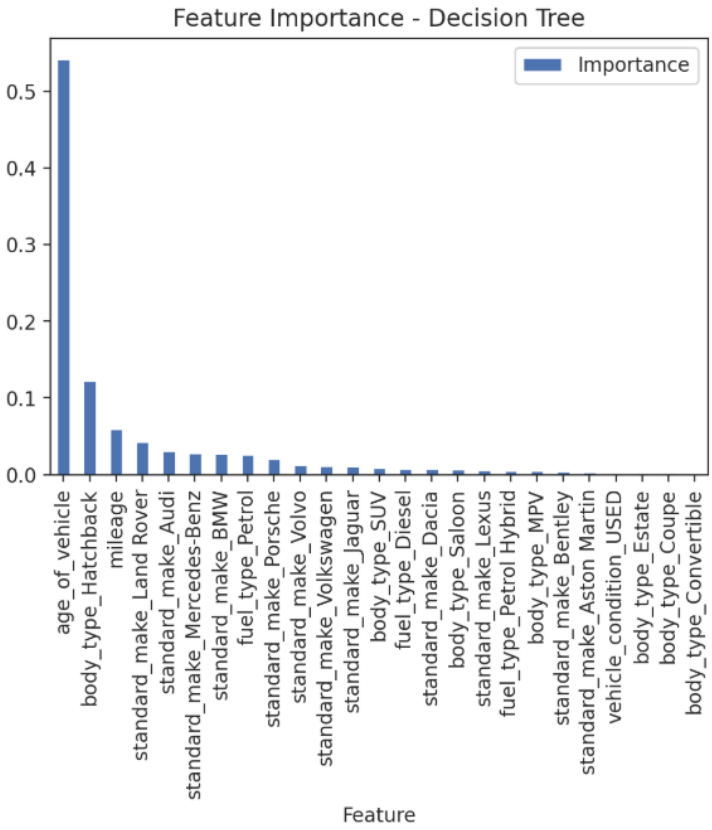


The actual and predicted graphs of kNN, decision tree and linear regression models are depicted in the figure that clearly depicts the efficiency of the models. The majorities of the points are near to the red dash line in the kNN plot, which is showing good predictive accuracy. This means that the basic structures in the data are well captured in kNN. The decision tree plot also shows that the model is fairly reasonable when aligned, but there is more scatter, which is more pronounced in extreme values. But the linear regression plot has the largest deviations from the red line especially as the actual values rise. This proves that the linear regression is not a good learner of the data since it only assumes that the data is linear and has a weak generalisation.



The prediction distribution plots highlight several differences between the models' performance. The kNN model shows a normal distribution of curve with shifting to the right, and highly peak values, which indicates a good generalisation and the tolerance to noises. The decision tree captures similar patterns but shows spikes due to its discrete nature. Linear regression provides a centred bell curve that can be extended far beyond zero, it is an indication that the model is inadequate when dealing with non-linear data correlation.

4.2 Feature Importance



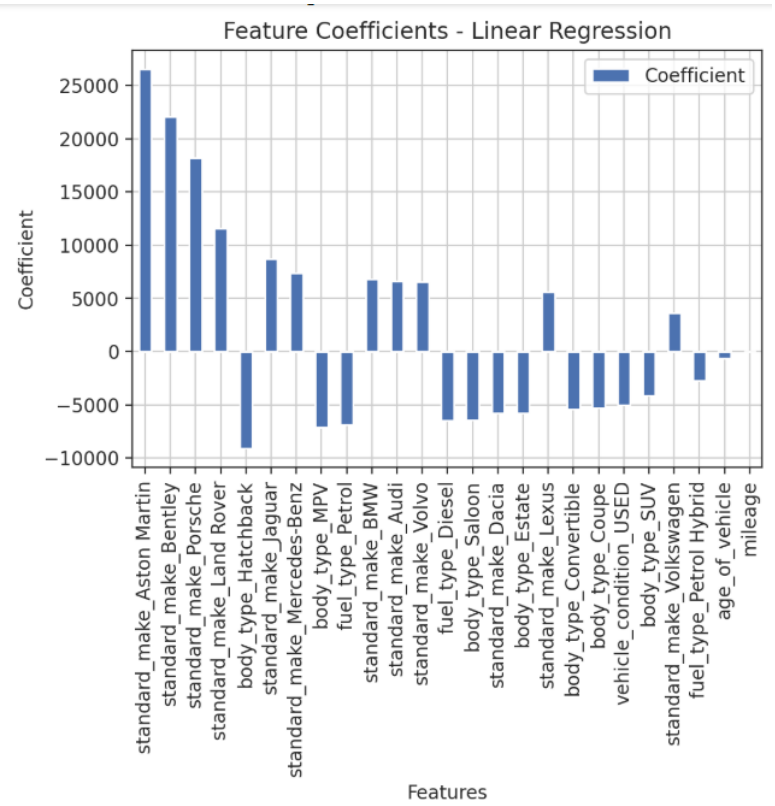
The image illustrates the feature used to predict the price in decision tree model. Age_of_vehicle show higher importance score to affirm that it is the most important feature that affects the price. This is in line with real world because value of vehicles decreases as the age of the vehicle increases. The second most important variable is body type (Hatchback) shows the impact of vehicle class on the rates. Because hatchbacks are among the most popular and versatile vehicles for many customers around the world, they are especially important for resale.

The other factor that is taken into consideration is referred to as the mileage. High mileage simply means that the car is highly used and therefore costs less value relative to a low mileage car which cost a lot since it is less used. Among the car manufacturers, the ones that exert considerable influence on prices include Land Rover and Audi, because the Tag on such brands is high because many people prefer quality car brands. Fuel type also meditates the analysis because the awareness of electric vehicles in recent years has grown due to the change in customer preferences and the need to save the environment.

Other features, such as body types (Coupe, Convertible) do not influence the prediction results in any way. These features may in fact have low variance or are less appearing in the dataset which may decrease their impact on the determination of price. Similarly, factors such as condition of the vehicle have a weak influence, probably because most of them fall under ordinary types of automobiles that are not significantly different in costs.

The feature importance distribution also shows that decision tree model uses only relevant features like age, mileage, and body type. The following conclusions support the model's interpretability because its decisions are consistent with conventional factors that affect vehicle prices.

	Feature	Coefficient
13	standard_make_Aston Martin	26552.701546
16	standard_make_Bentley	22087.160556
22	standard_make_Porsche	18174.224477
19	standard_make_Land Rover	11606.502210
8	body_type_Hatchback	-9063.806720
18	standard_make_Jaguar	8739.403447
21	standard_make_Mercedes-Benz	7389.076285
9	body_type_MPV	-7117.069923
3	fuel_type_Petrol	-6887.033860
15	standard_make_BMW	6781.734462
14	standard_make_Audi	6631.142772
24	standard_make_Volvo	6568.167596
2	fuel_type_Diesel	-6474.115025
11	body_type_Saloon	-6438.306969
17	standard_make_Dacia	-5800.037864
7	body_type_Estate	-5763.227958
20	standard_make_Lexus	5620.936082
5	body_type_Convertible	-5432.208240
6	body_type_Coupe	-5243.729895
12	vehicle_condition_USED	-5050.753074
10	body_type_SUV	-4101.477648
23	standard_make_Volkswagen	3657.172485
4	fuel_type_Petrol Hybrid	-2735.168829
1	age_of_vehicle	-611.766954
0	mileage	-0.083319



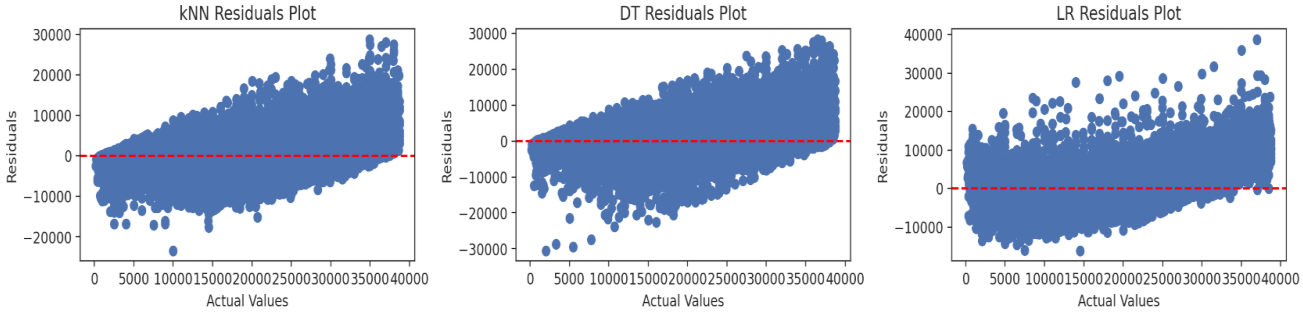
From the coefficients graph of linear regression, standard_make (Aston Martin, Bentley, Porsche) have the highest positive coefficients showing that the inclusion of the premium brands increases the cost of vehicles. This can be related with the actual trends as these brands are associated with prestige and high quality. Also, standard_make (Land Rover, Mercedes-Benz) has a positive coefficient with price, and this highlights the impact that brand has on the pricing of the vehicle.

On the other hand, the coefficients of body_type (Hatchback) as well as mileage are negative. The predicted price decreases as compared to other body types because hatchback is a more economical car type. The car's resale value decreases when it's used for more miles. Age_of_vehicle has almost no effect and this may be due to some extent to overlap or multicollinearity with other variables such as mileage.

Fuel type like petrol and diesel and other body types such as coupe, convertible shows a moderate impact on the final price. Standard_make (Volkswagen) also experience minor impacts, or lower variability or correlation with the price in this data set as seen in the actual value of USED for vehicle_condition.

The distribution of coefficients shows that luxury brands and premium options increasing prices, while such features as mileage and body type (Hatchback) decreasing them. This analysis also shows that Linear Regression is easy to interpret, giving direct information on how each feature influences the prediction of vehicle prices. But it also means that the model does not optimally capture nonlinear relationships as it pertains to some features such as age and mileage.

4.3 Fine-Grained Evaluation



In the kNN residual plot, it is observed that the residuals are fairly well centred around the zero line for most actual values though there is an upward trend for the higher actual values. This means that though kNN provides good results overall, it gives slightly lower estimates of price for relatively costly cars.

The decision tree residual plot shows, with most of the residuals located near zero for small and average actual values. But for greater actual values, there is more scatter and the residuals appear to be more random. This suggests that decision tree model may not generalise well when predicting prices of car above certain value.

The linear regression residual plot shows the greatest variability of residuals and even a stronger pattern. From the residuals chart, linear regression underestimates the value when actual value was high due to its inability to capture non-linear relationship. Further, it is evident that the residuals are spread out across all value ranges with greater variability than kNN and DT.

	Actual Value	kNN Predicted	kNN Residual	Decision Tree Predicted	Decision Tree Residual	Linear Regression Predicted	Linear Regression Residual
120176	8999	13873.363636	-4874.363636	11619.561285	-2620.561285	14220.197512	-5221.197512
105644	15499	16394.000000	-895.000000	17061.803030	-1562.803030	17436.059658	-1937.059658
180345	13400	15479.818182	-2079.818182	16600.000000	-3200.000000	16182.309642	-2782.309642
208691	9798	10756.454545	-958.454545	11719.506438	-1921.506438	12093.762393	-2295.762393
156103	6495	5995.181818	499.818182	6518.584906	-23.584906	6234.535683	260.464317

The table compares instance-level predictions and residuals for the kNN, decision tree, and linear regression models. kNN shows relatively low residuals, demonstrating consistent and accurate predictions, though it slightly overestimates in some cases (e.g., instance 120176, residual: 4,874.36). The decision tree also follows the trend well, but has slightly higher residuals such as the one in 180345 (3,200.00). Linear regression shows the largest residuals, reflecting its limitations in capturing non-linear relationships (e.g., instance 120176, residual: 5,221.19).

Reference

https://en.wikipedia.org/wiki/Vehicle_registration_plates_of_the_United_Kingdom