DS Case Study
Toyota Used Car Sales

Author: Lam Trinh

Table of contents

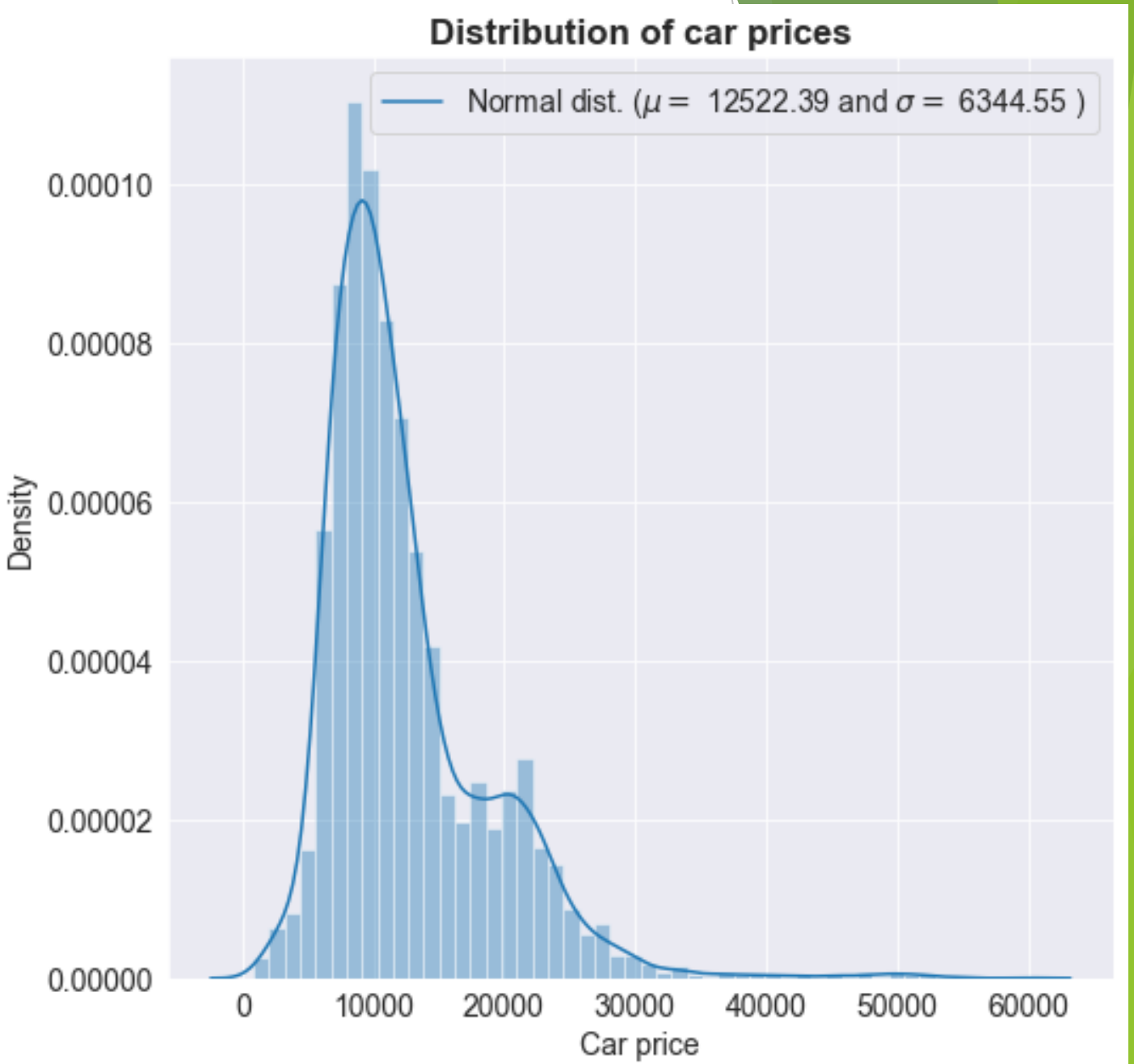
Data overview

- There are 6,738 data points of cars. There's no missing (or null) values in the data.
- The range of year for cars is from 1998 to 2020, with most of the cars belong to the model year 2016
- Prices range from under 1,000 pounds all the way to almost 60,000 pounds
- Mileage ranges from brand-new (with just 2 miles on odometer) to 174,000 miles.
- There're 18 unique models. All are Toyota brand.

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
count	6738	6738.000000	6738.000000	6738	6738.000000	6738	6738.000000	6738.000000	6738.000000
unique	18	NaN	NaN	4	NaN	4	NaN	NaN	NaN
top	Yaris	NaN	NaN	Manual	NaN	Petrol	NaN	NaN	NaN
freq	2122	NaN	NaN	3826	NaN	4087	NaN	NaN	NaN
mean	NaN	2016.748145	12522.391066	NaN	22857.413921	NaN	94.697240	63.042223	1.471297
std	NaN	2.204062	6345.017587	NaN	19125.464147	NaN	73.880776	15.836710	0.436159
min	NaN	1998.000000	850.000000	NaN	2.000000	NaN	0.000000	2.800000	0.000000
25%	NaN	2016.000000	8290.000000	NaN	9446.000000	NaN	0.000000	55.400000	1.000000
50%	NaN	2017.000000	10795.000000	NaN	18513.000000	NaN	135.000000	62.800000	1.500000
75%	NaN	2018.000000	14995.000000	NaN	31063.750000	NaN	145.000000	69.000000	1.800000
max	NaN	2020.000000	59995.000000	NaN	174419.000000	NaN	565.000000	235.000000	4.500000

Visualization (1)

- Most of the prices seem to congregate around the 8,000-15,000 range, with a mean price of 12,500 pounds.
- From the graphs, we can also see that the distribution of prices is right-skewed. That is, the mean (mu) or average price is larger than the median (the middle price point that divides half of the data set).

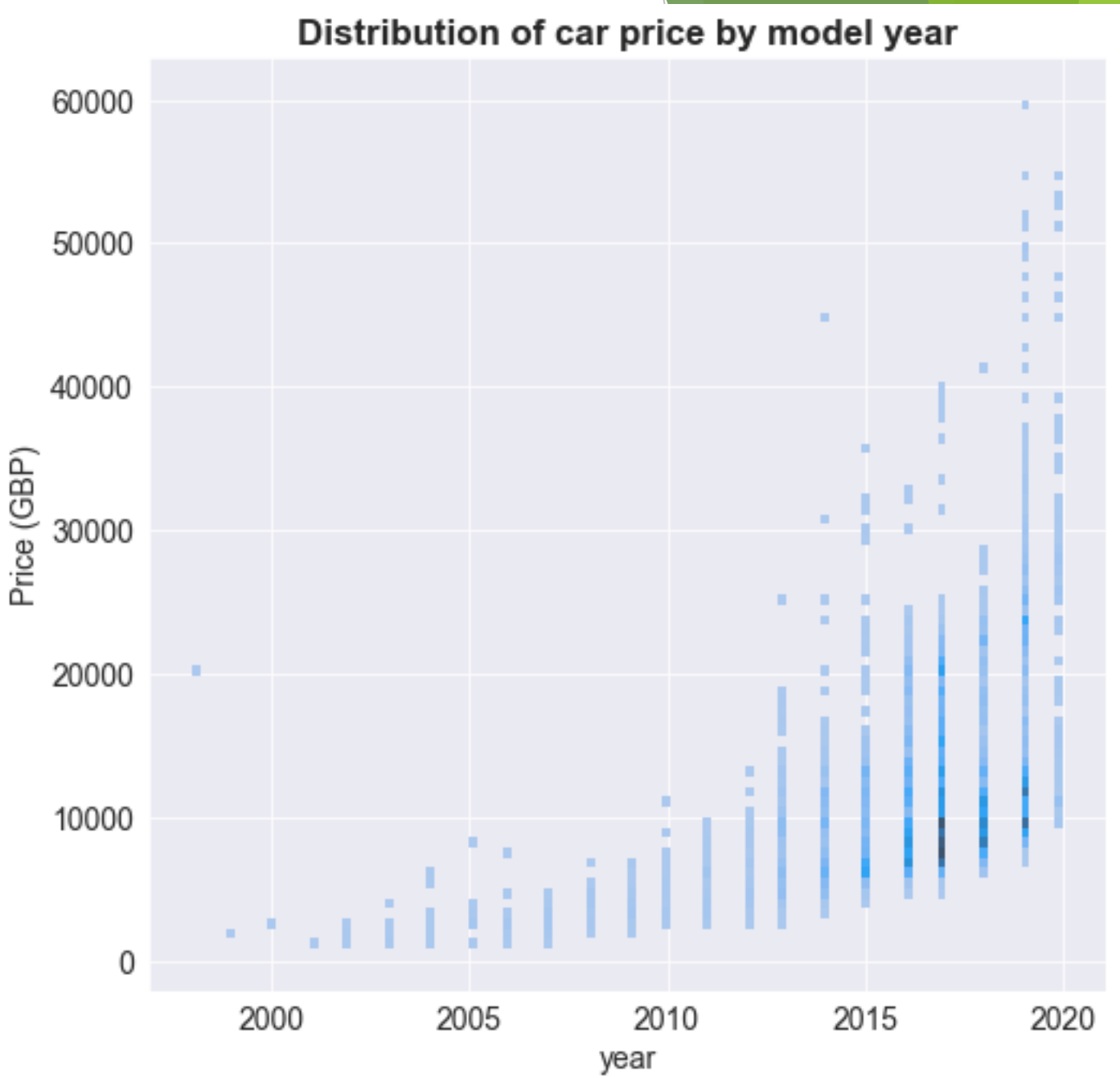


Visualization (2)

- The median car price for manual seems to be around 10,000 GBP, whereas for automatic it's 1.5 times more expensive, at 15,000
- Also, according to this plot, we may have some outliers (points that are larger than 3rd quartile + 1.5 times interquartile range) for cars with automatic transmission.

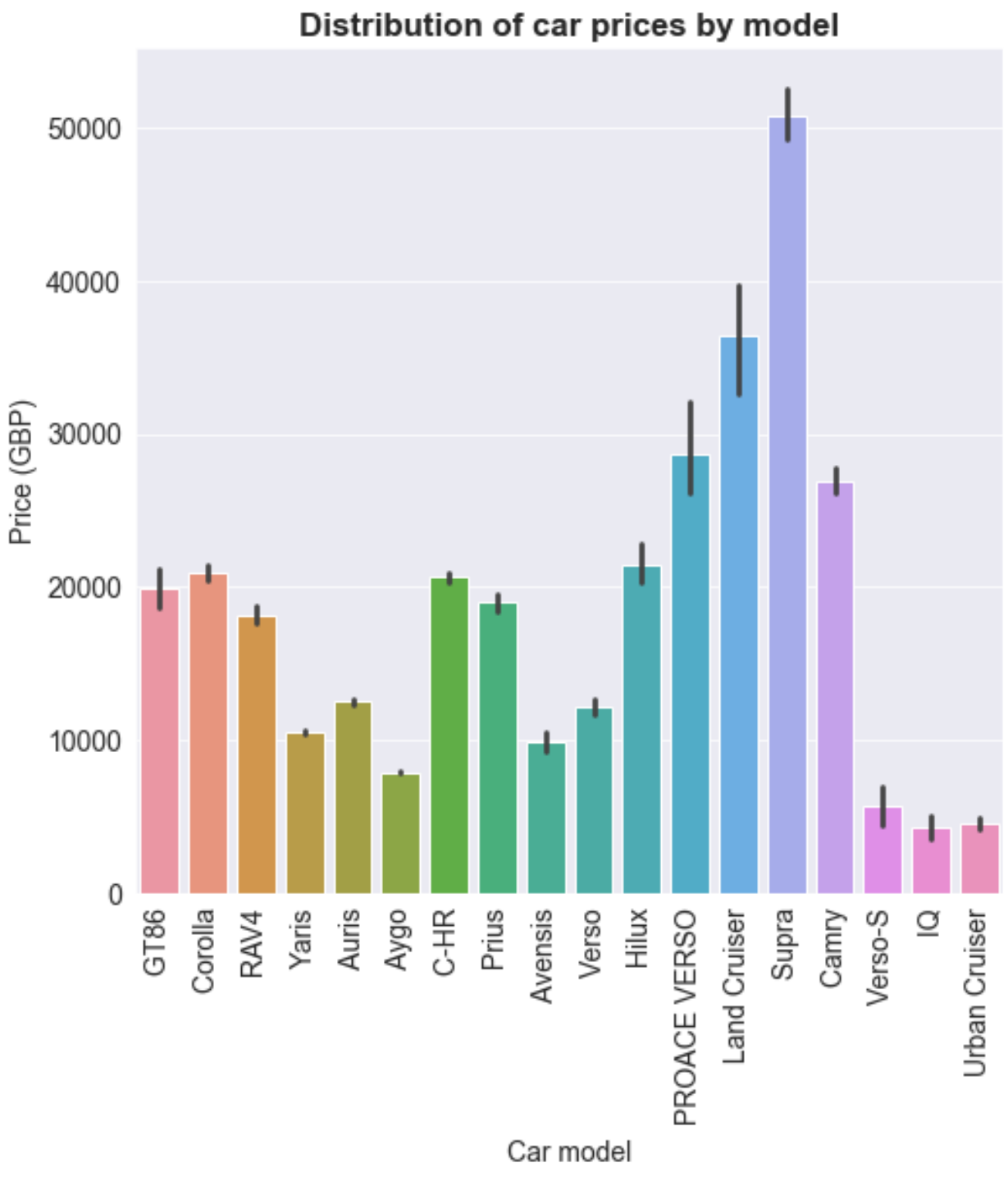
Visualization (3)

- In general, the newer a car is, the more price it will command in the market
- We have more cars that are less than 10 year old than cars that are older than 10 year old.



Visualization (4)

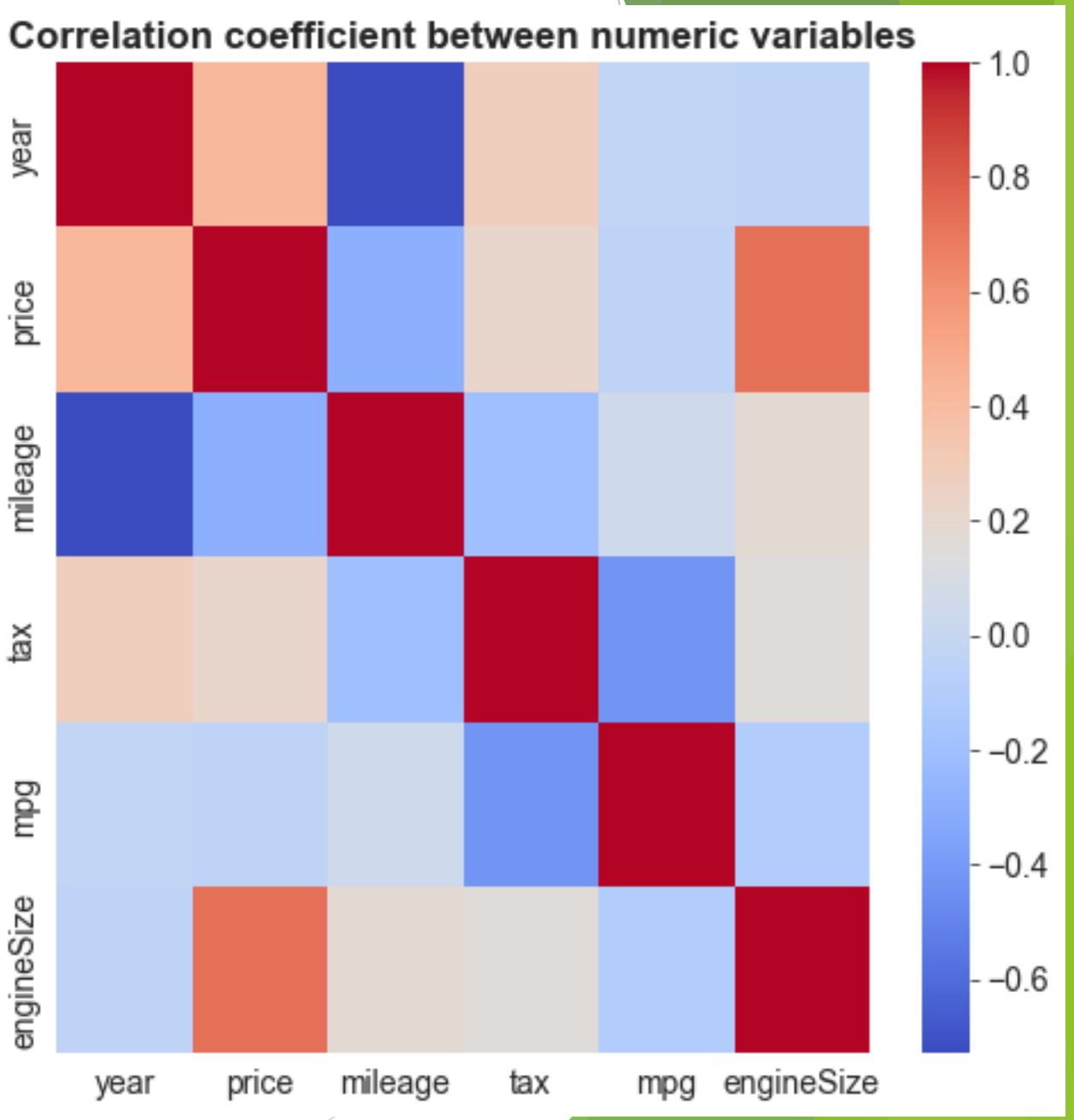
- Supra and Land Cruiser seem to the two most expensive models at about 50,000 and 36,000 GBP respectively.
- The most affordable models are Urban Cruiser and IQ at 4,600 and 4,200 GBP respectively.



Visualization (5)

From the heatmap, we can see that:

- Price has a strong positive correlation with model year. As model year increases (cars are newer), price will increase.
- Price has a strong negative correlation with mileage, which makes sense. As mileage goes up, price will go down.
- Mileage has strong negative correlation with model year.
- The most surprising finding is price has a strong positive correlation with size of the engine. As the size of engine increases, price will increase.



Methodology: for predicting car prices

- Pre-process the data so that it can be modeled using machine learning.
- Build 4 different models and measure the score of these models to figure out which model has the best score on a validation dataset (dataset that is given). The models are listed from most simple to more complex:
  - Decision Tree Regressor
  - Random Forest Regressor
  - AdaBoost Regressor
  - Gradient Boosting Regressor
- Choose the model that has the best score from the validation dataset (after hyperparameters tuning).
- From there, train the model and use it to predict prices of cars on new data (test set).
- Calculate to see what percentage of the predicted prices do not exceed 1,500 GBP of estimated prices.
- Report the score and the results.

Preprocessing

- Created dummy variables to for each of the categorical variables: model, transmission, and fuelType.
- New data has 28 independent variables, the dependent variable is price.
- We split our data into 3 parts: 60% is training data, 20% is validation data (for validating and choosing the best model), and 20% is test data (for final model evaluation).

Preprocessing

Since we have a few categorical variables, we'll need to create dummy variables for them in order to incorporate them into the model

```
Numeric = ['year', 'price','mileage','tax','mpg','engineSize']
Category = ['model', 'transmission', 'fuelType']

# Create dummy variables for the categorical variables, dropping 1 variable out so that they will not related to one another
new_df = pd.get_dummies(df, columns = Category, drop_first = True)
print(new_df.columns)
```

```
Index(['year', 'price', 'mileage', 'tax', 'mpg', 'engineSize',
       'model_ Avensis', 'model_ Aygo', 'model_ C-HR', 'model_ Camry',
       'model_ Corolla', 'model_ GT86', 'model_ Hilux', 'model_ IQ',
       'model_ Land Cruiser', 'model_ PROACE VERSO', 'model_ Prius',
       'model_ RAV4', 'model_ Supra', 'model_ Urban Cruiser', 'model_ Verso',
       'model_ Verso-S', 'model_ Yaris', 'transmission_Manual',
       'transmission_Other', 'transmission_Semi-Auto', 'fuelType_Hybrid',
       'fuelType_Other', 'fuelType_Petrol'],
      dtype='object')
```

Model validation and selection

- The results shown are for validation data (data that we presumably have seen).
- Based on the result shown in the tables, we can see that Ada Boost Regressor has the best score Root Mean Square Error (RMSE) score, best Mean Absolute Error (MAE) score, and also the best fit R^2 score. So this will be the model that we will use on the test set to predict prices of cars.
- Basically, Mean Absolute Error means the absolute difference between the predicted price and the actual estimated price. A smaller number would be better in this case as it indicates that the predicted price is more aligned with the estimated price.
- R^2 score signifies the percentage of price that can be explained by the model, it ranges between 0 and 1, and the closer to 1, the better the model is.

Model used	RMSE score
Decision Tree Regressor	1550.9
Random Forest Regressor	1639.2
Ada Boost Regressor	1378.4
Gradient Boosting Regressor	1746.2

Model used	Mean Absolute Error score
Decision Tree Regressor	964.4
Random Forest Regressor	990.3
Ada Boost Regressor	865.9
Gradient Boosting Regressor	1165.4

Model used	R^2 score
Decision Tree Regressor	93.6%
Random Forest Regressor	92.8%
Ada Boost Regressor	94.9%
Gradient Boosting Regressor	91.9%

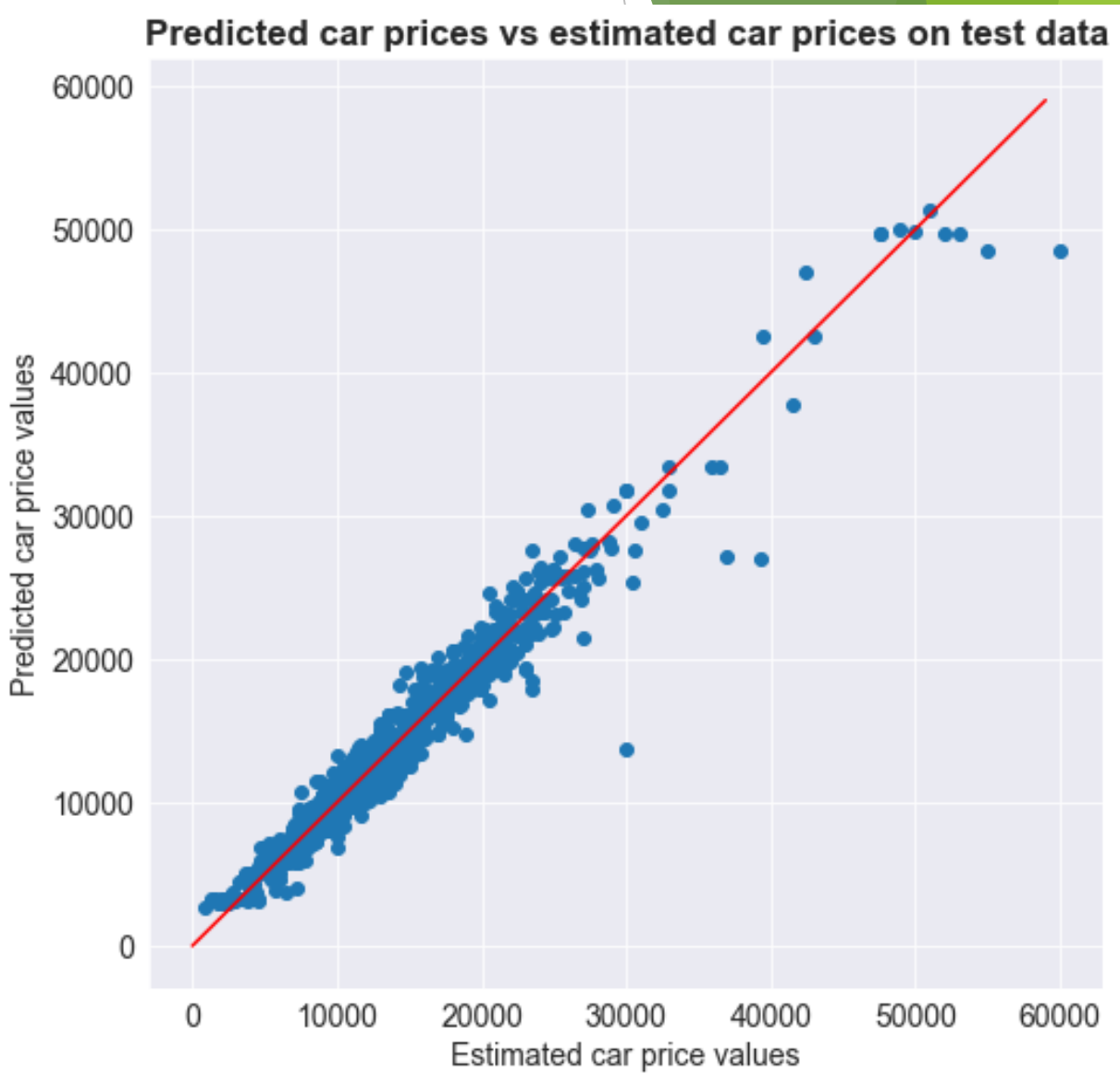
Final model evaluation (1)

- Now we will show the result of the prediction from the best model: Ada Boost Regressor.
- We will use this model to predict the test data (data that our models have not seen yet).
- We obtain a final mean absolute error score of 837, meaning that the difference between predicted car price and estimated price is about 837 pounds.

Ada Boost Regressor evaluation method	Score
Root mean square error	1311.4
Mean absolute error	837.9
R^2 score	96.2%

Final model evaluation (2)

- X (horizontal) axis represents estimated car price.
- Y (vertical) axis represents predicted car price.
- The further away the points are from the red line, the worse the fit.
- Points under the red line signify under-predicting (predicted price < estimated price), whereas points above the red line signify over-predicting (predicted price > estimated price)
- We can see that our predicted price is quite good, as there're very few predicted price points that deviate too much from the estimated price points.



Final model evaluation (3)

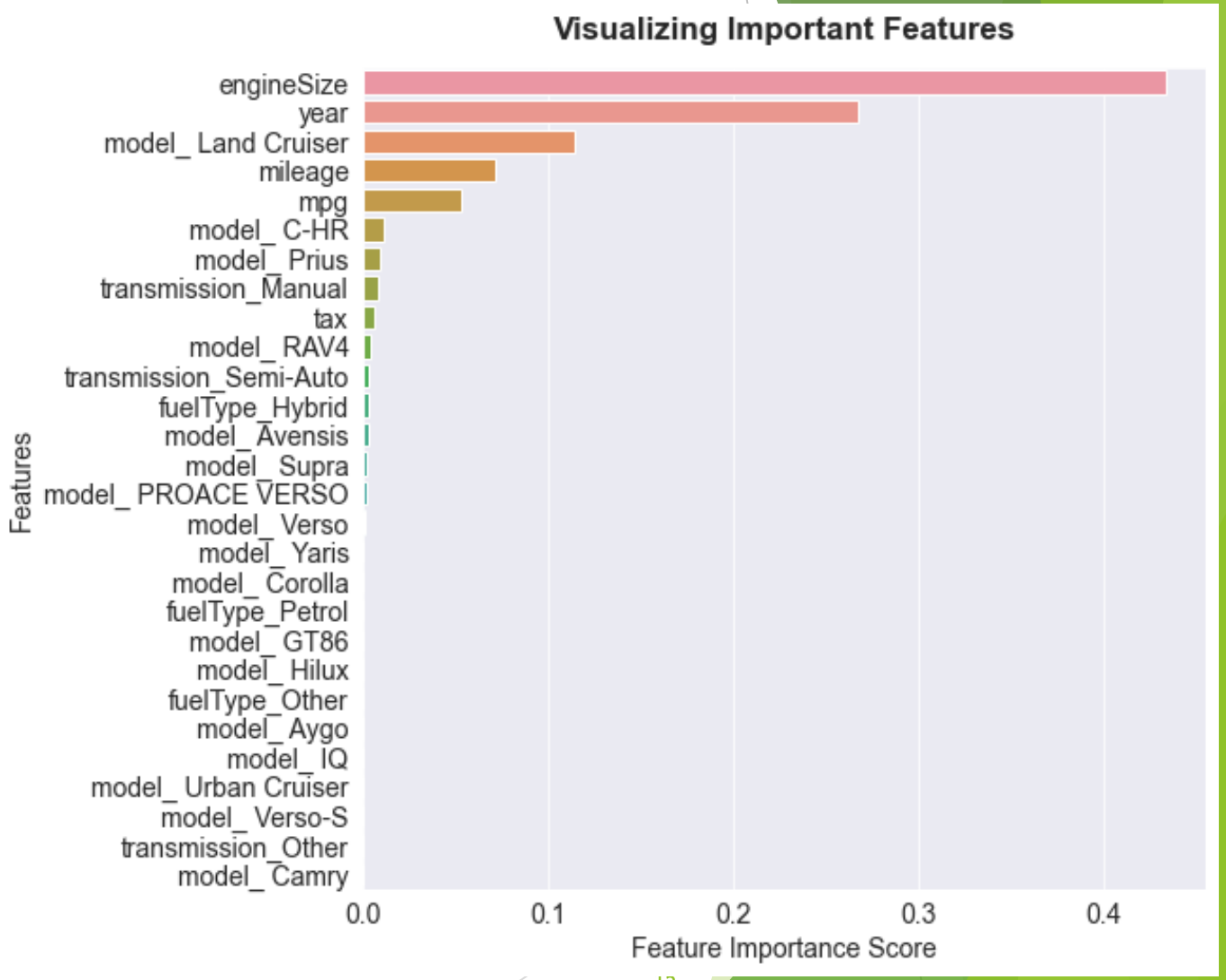
- This is a business metric: accuracy of the model given business criteria.
- If we only consider predicted prices that are 1,500 GBP more than the estimated car prices as failures, then we can calculate the model's accuracy score.
- Logic: if the predicted price is 1,500 GBP less than the estimated price, we can still sell the car, even though it might end up being at a loss, depending on the cost of the car. Since we're dealing with used cars, there is typically more margin than new cars.
- Out of 1,348 data points in the set, there are only 100 price points where the predicted values are 1,500 GBP more than estimated car prices.
- So accuracy of our model is (1348 – 100)/1348 = **92.58%**

Feature importance

- The top 5 features that are most predictive of a car's prices are, in order of importance:

1. Engine size
2. Model year
3. Whether the car is a Land Cruiser
4. Mileage on the car
5. Miles per gallon (fuel economy)

- The rest of the features do not have much power in predicting a used car's price in this model.



Final remarks

- I attempted to build another model with just the 5 or even 9 most-important features, but it didn't improve the score of the model, so those results are not presented here.
- We can compare the results of the predictions here with other industry sources (e.g., Kelly Blue Book) to reference the sales price.
- We may be able to deduce further price information from the car if we know its overall condition: "Like new", "Very good", "Good", "Okay", "Need Repair."