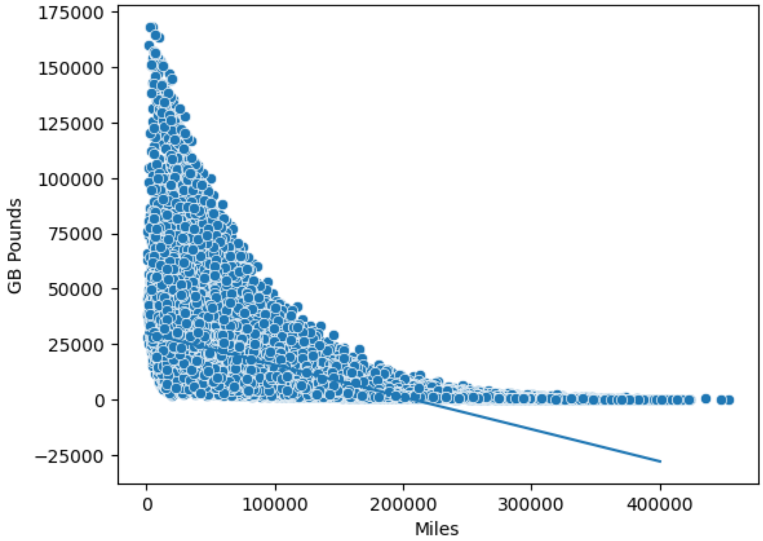
1. Introduction

Predictions can be made in machine learning using both supervised and unsupervised learning. Regression analyses are supervised learning methods for finding the best trend line to describe pattern in a data (Theobald, 2017). Neural networks like CNN (is unsupervised) work better at prediction as they learn the way humans do (expressanalytics.com, 2024). Clustering fall under both supervised and unsupervised (Theobald, 2017). I shall experiment on predictions of car price using regression models, CNN and clustering algorithm.

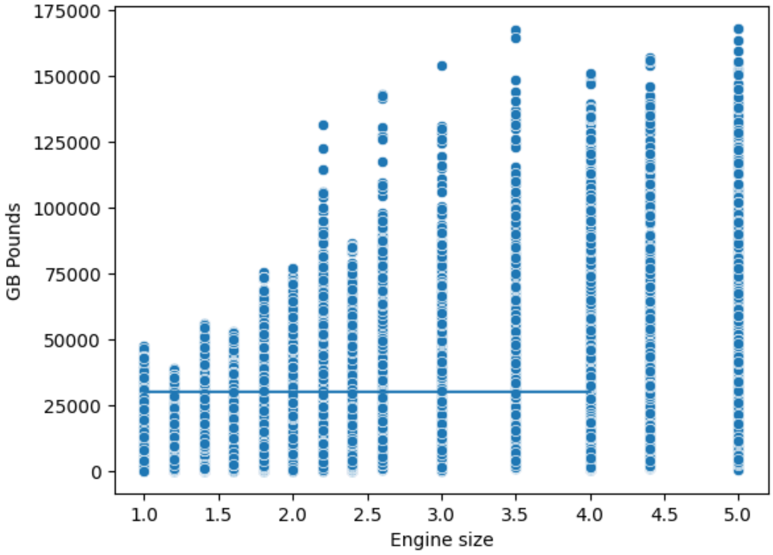
Simple Linear Regression Model Using the Numeric Columns

1. Price v Mileage



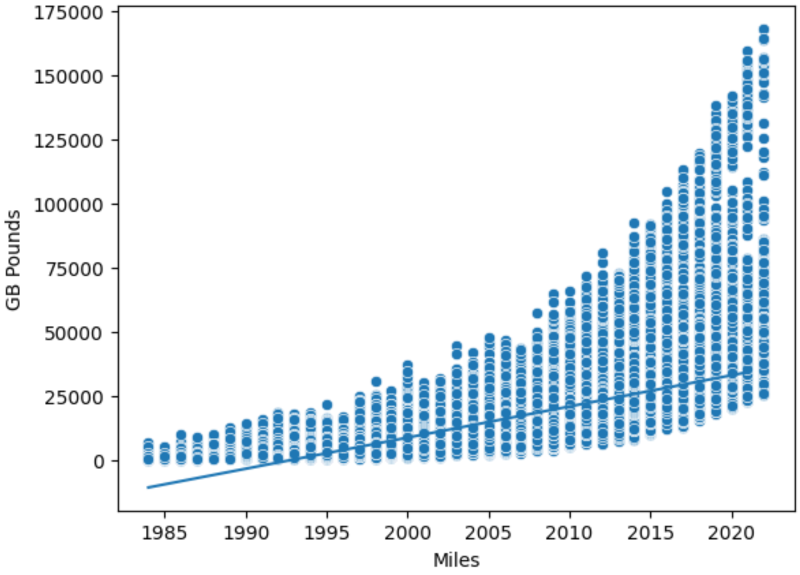
Mean squared error: 162468566.87254104
R2: 0.4013139100884707

2. Prive v Engine size



Mean squared error: 230499154.45279127
R2: 0.15062562461380213

3. Price v Year of manufacture

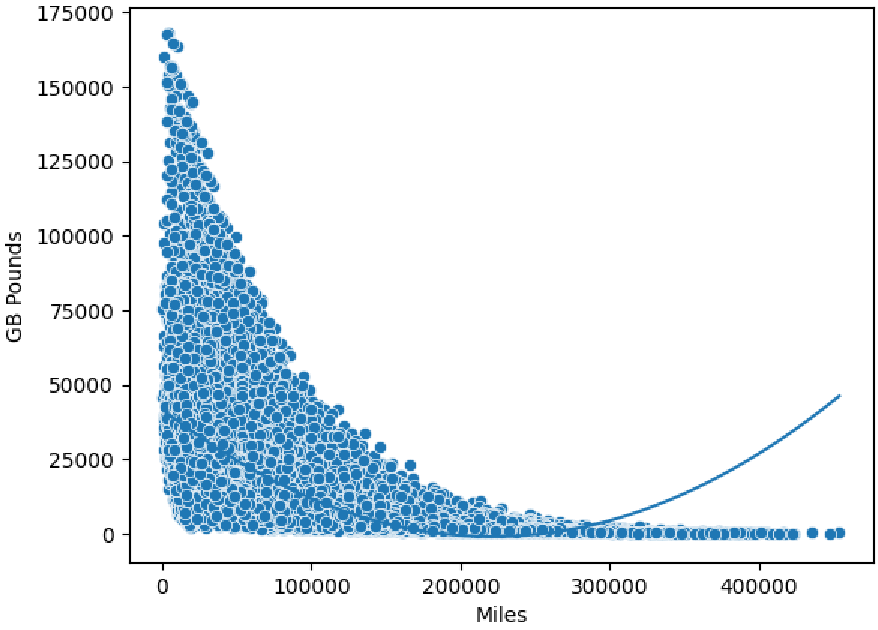


Mean squared error: 132678999.94793083
R2: 0.5110865244812856

The predicted line grazes below most data points. 'Year of manufacture is best predictor with R2 score of .51 and MSE of 132678999.95

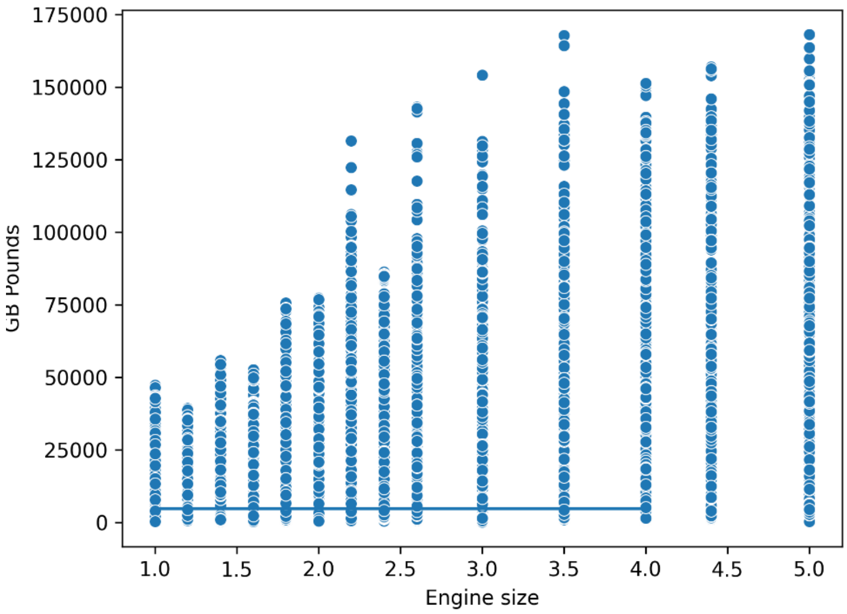
Simple Polynomial Regression Model Using the Numeric Columns

1. Price v Mileage



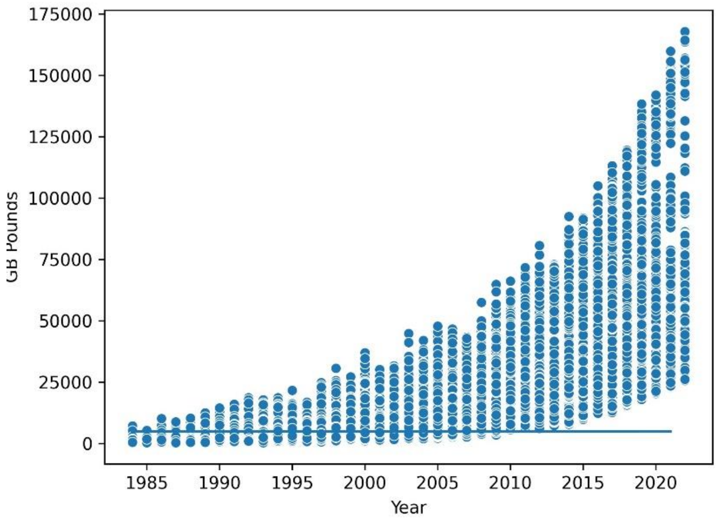
Mean squared error: 129620312.1626197
R2: 0.5223575898060919

2. Price v Engine size



Mean squared error: 230326165.99946904
R2: 0.15126307580028653

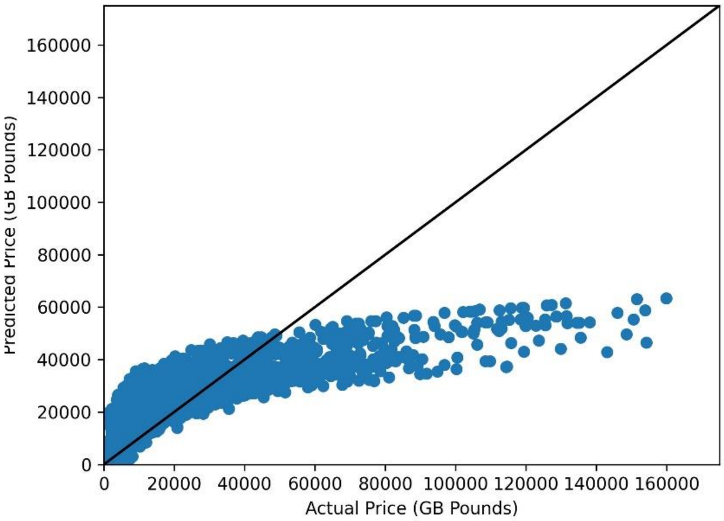
3. Price v Year of manufacture



Mean squared error: 105993894.20194323
R2: 0.6094194015754401

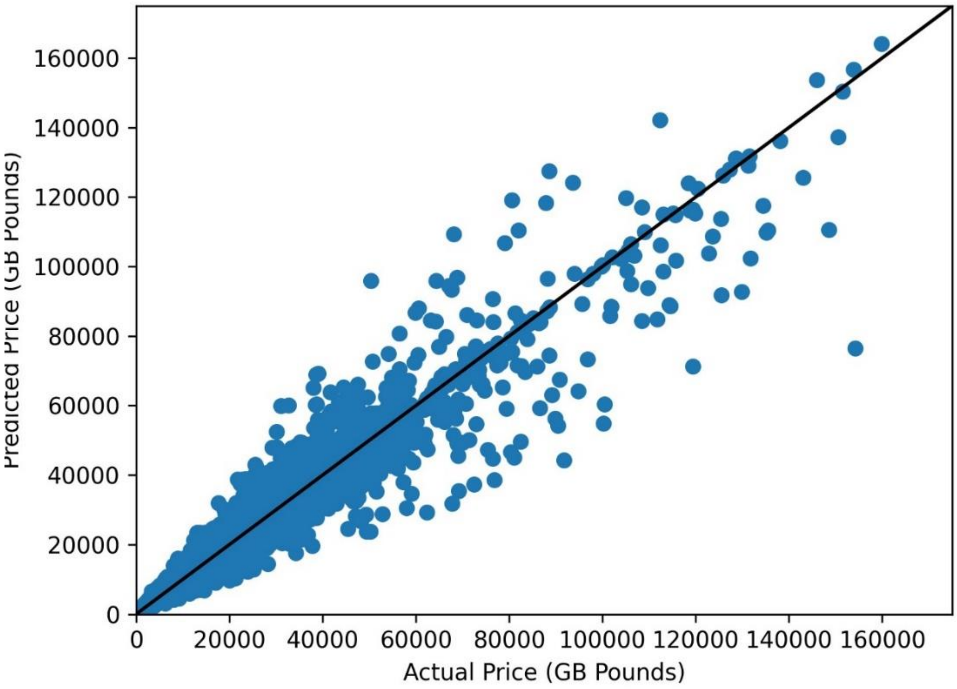
Answer 1a(i): Comparing the above plots and computation, Year of Manufacture is a better predictor as in simple and polynomial regression above, it gave a higher R2 score .61 and lesser MSE 105,993,894.2 and, therefore, the better prediction (Palmer, 2009). **1a(i)**: In both simple and polynomial, the predicted line did not fit well. However, Price v Mileage did better in both simple and polynomial with the polynomial fitting better. In Price v Year of manufacture as well as Price v Engine size, the predictor line of the linear cut through more data points than in polynomial.

1b Multiple Linear Regression with Numerical Variables



Mean squared error: 89158615.76017143
R2: 0.671456306417368

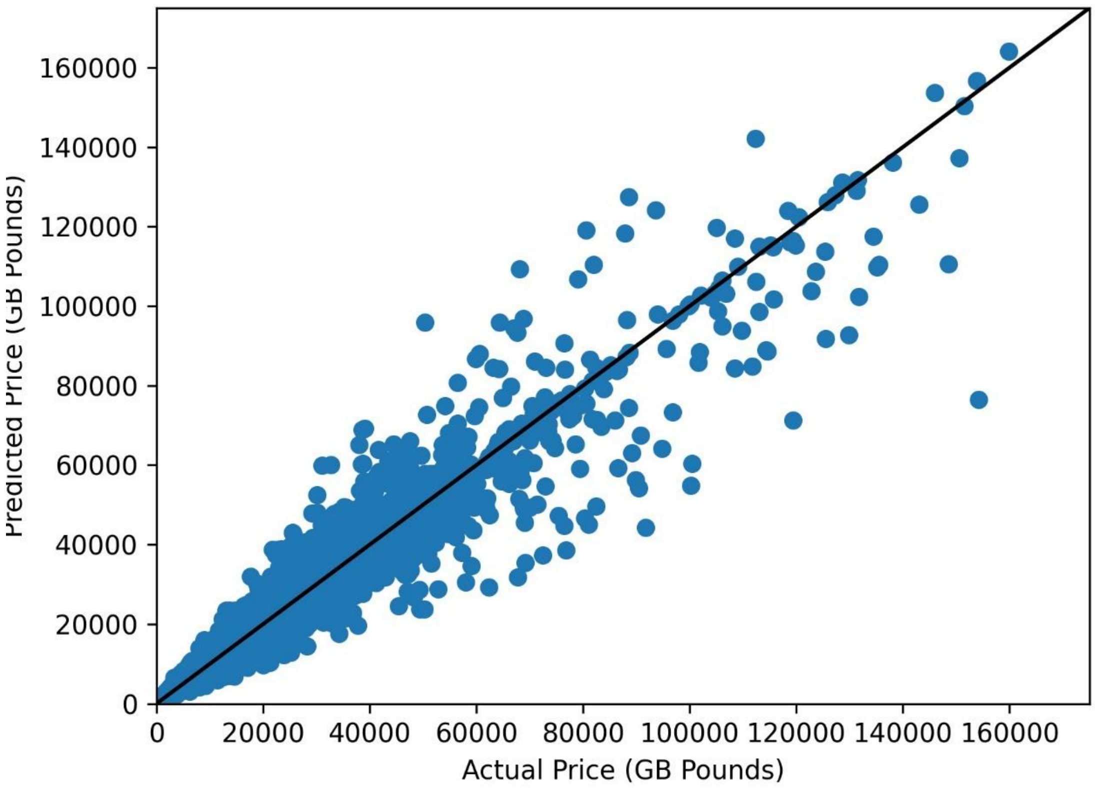
Random Forest with Numerical Variables



Mean squared error: 20159131.522860516
R2: 0.9257149129843565

Answer 1b: Yes, multiple input features improved the accuracy of the model's prediction as there is improvement in the R2 scores (.93) and reduction in error as seen in MSE (20,159,131.52)

1c. Random Forest with Numerical + Categorical Variables



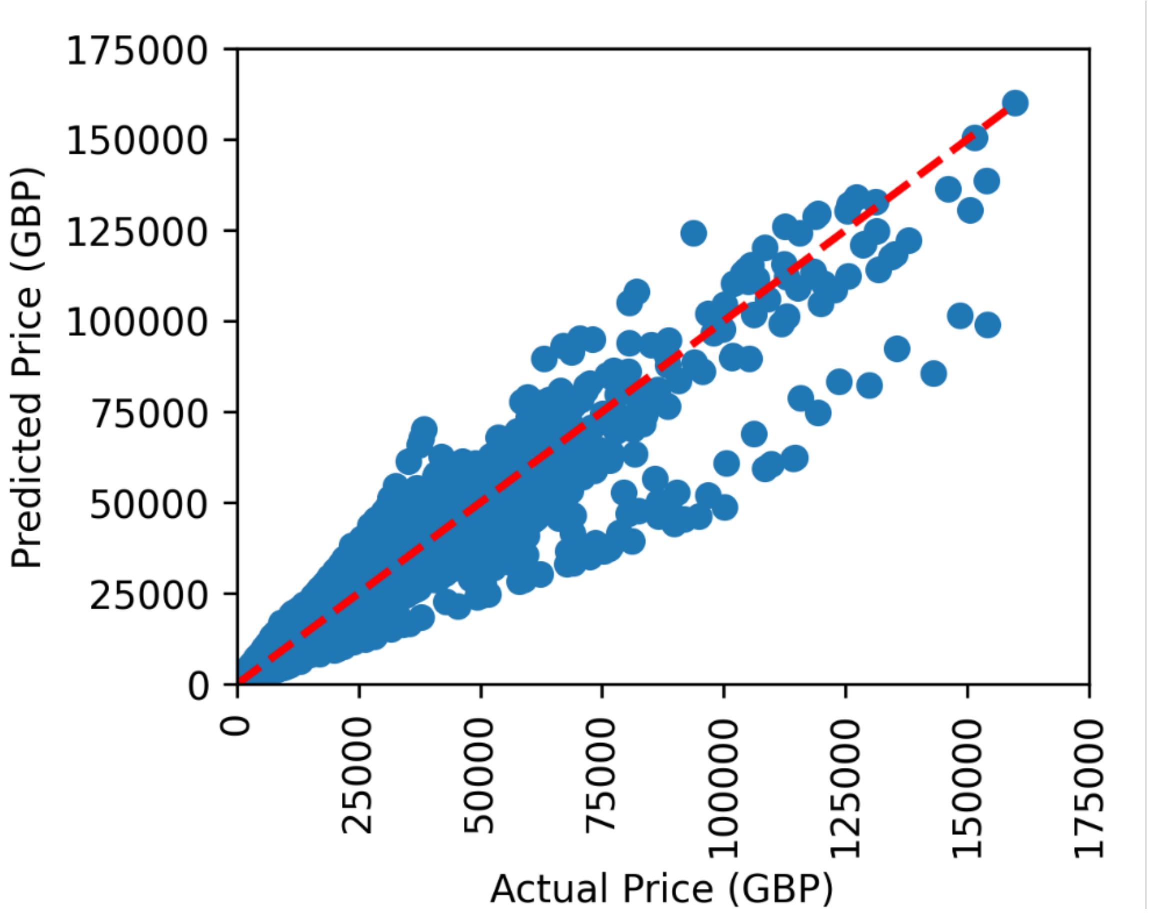
```
Mean squared error:  14529057.213402066
R2:  0.9464613702168213
```

With a near 1 R2 score of .95 recorded here and a reduced MSE of 14,529,057, we can conclude that random Forest model is the best compared to Linear and multiple regression with lower R2 score and high MSE rates.

Answer 1c: Yes, Random Forest Regressor with categorical and numerical variables improved the accuracy of the model with R2 score of .95 and a reduced MSE of 14,529,057. Random Forest model is the best compared to Linear and multiple regression with lower R2 score and high MSE rates (Igual and Segui, 2017).

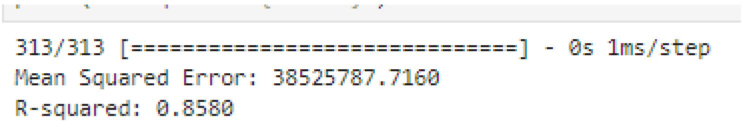
1d. ANN Model 1

ANN



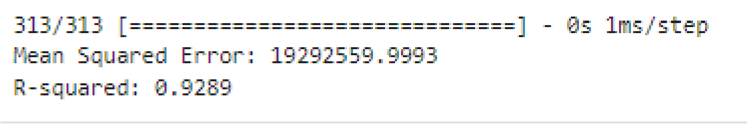
```
Mean Squared Error: 20055340.7003
R-squared: 0.9261
```

Models 2 : Reducing the learning rate



```
313/313 [==============================] - 0s 1ms/step
Mean Squared Error: 38525787.7160
R-squared: 0.8580
```

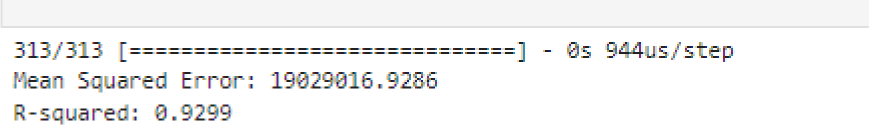
Model 3: Tuning the number of layers



```
313/313 [==============================] - 0s 1ms/step
Mean Squared Error: 19292559.9993
R-squared: 0.9289
```

ANN Model 4: Add drop out rate of 10%

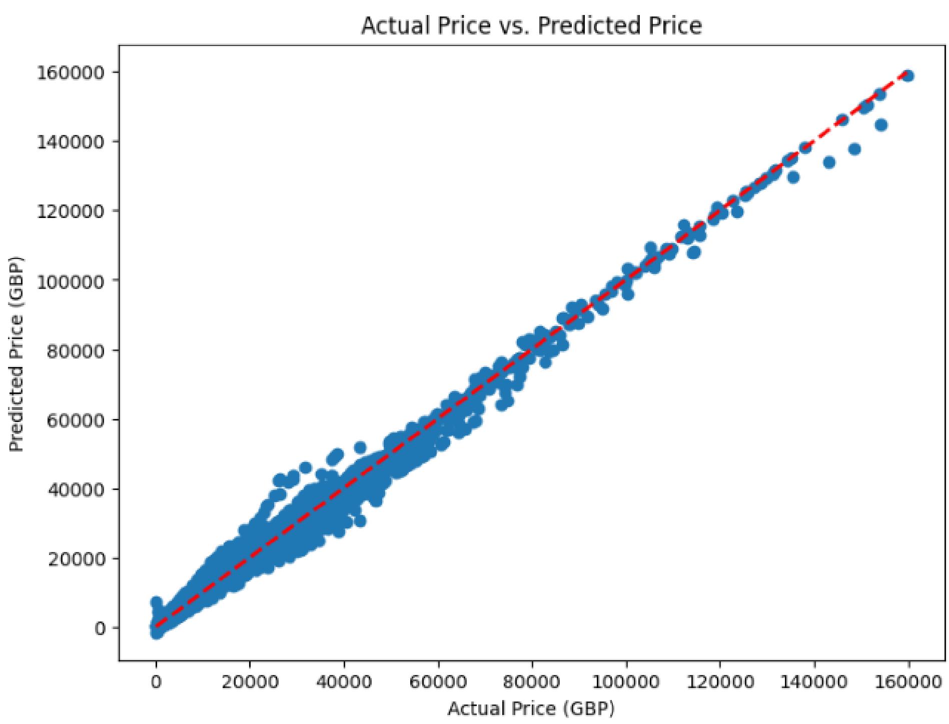
This first model improved



```
313/313 [==============================] - 0s 944us/step
Mean Squared Error: 19029016.9286
R-squared: 0.9299
```

ANN: Numerical and Numerical

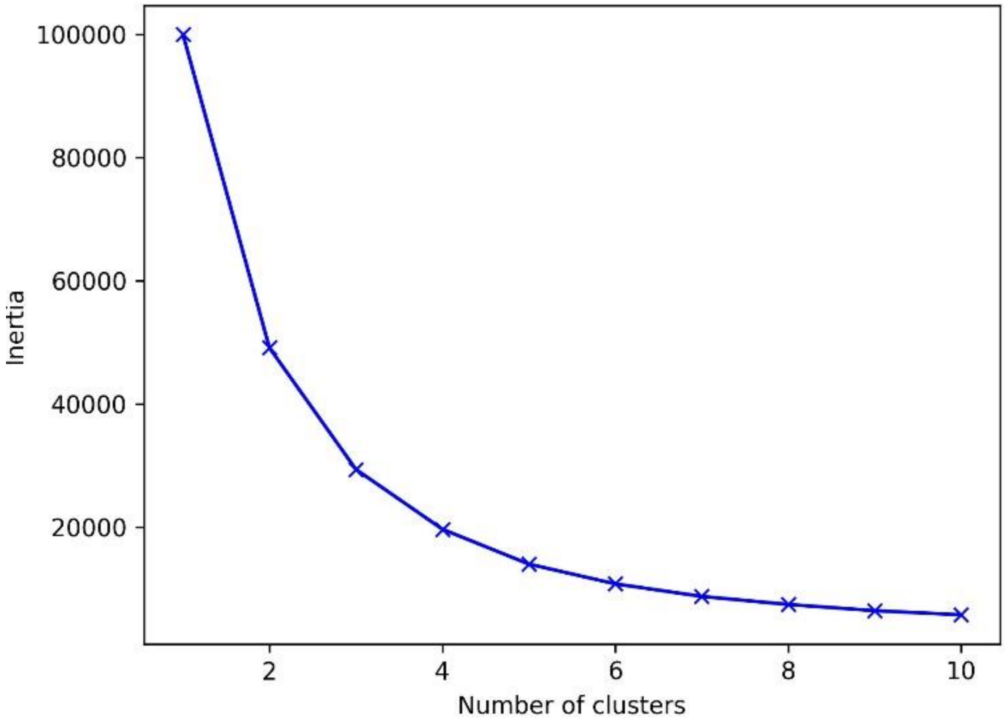
Mean Squared Error: 2198945.5261
R-squared: 0.9919



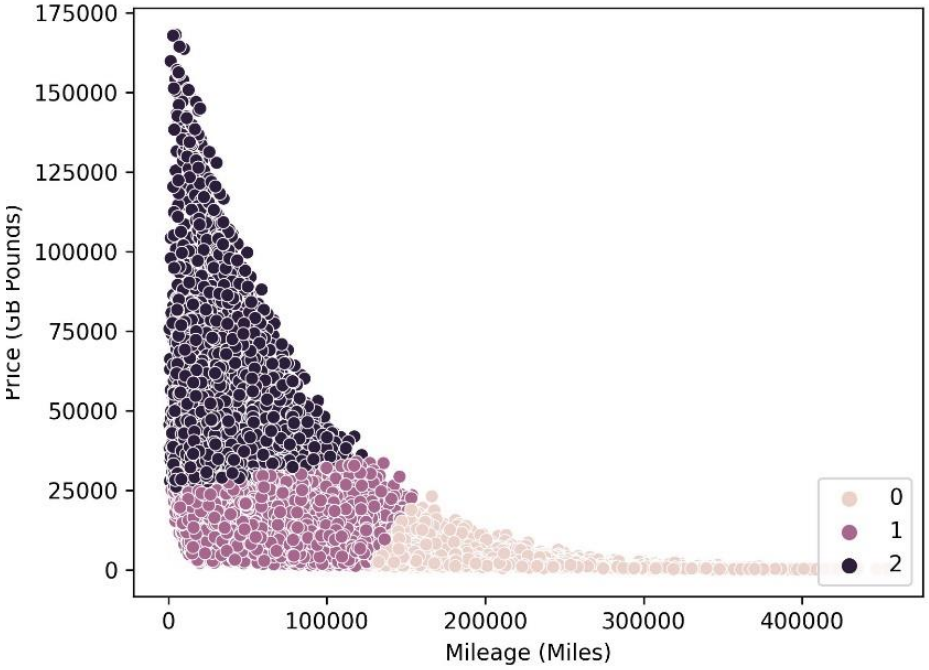
1e ANN with numerical and categorical variablest is the best model as its MSE error is lower and R2 score of .99 is near 1. R2 indicates how well a model predicts responses for new observations and is better compared to SMAPE, MAE, MAPE, MSE and RMSE (Chicco, 2021)

1f k-Means Clustering

Price v Mileage

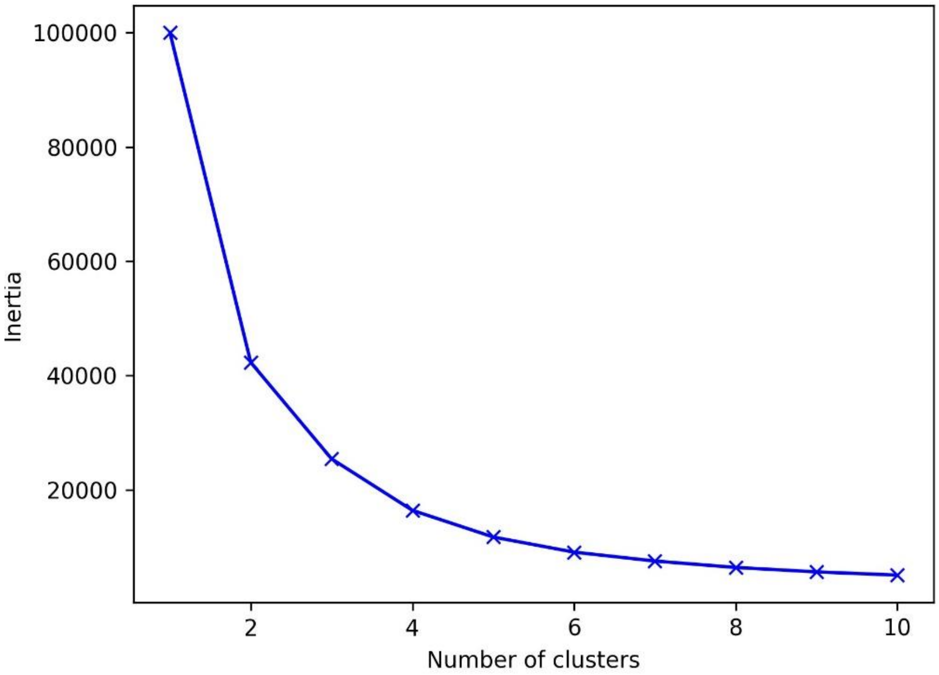


3 is the optimal number of clusters to use here as it is the point of the elbow.

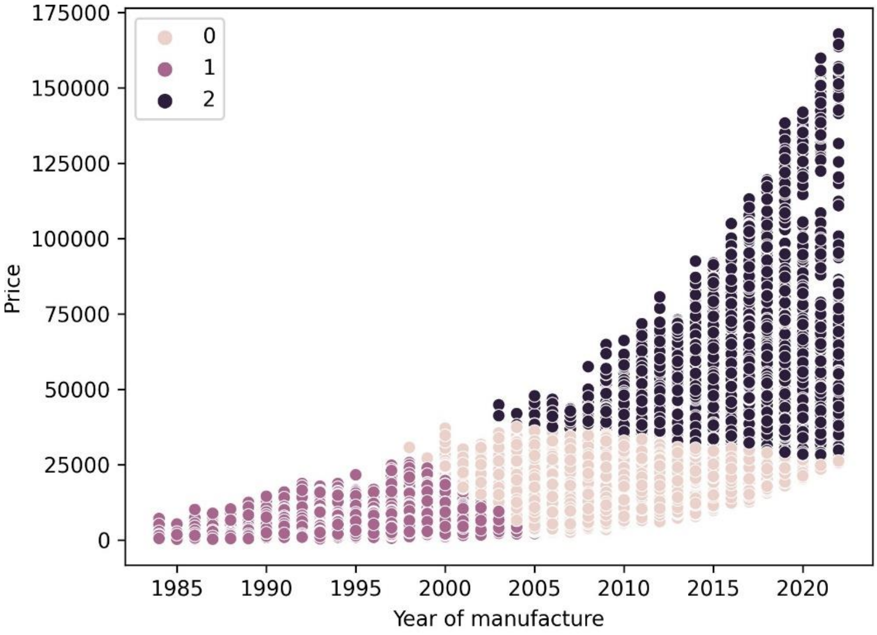


Davies Bouldin Index: 0.6870
Silhouette Coefficient: 0.4786

Price v Year of Manufacture

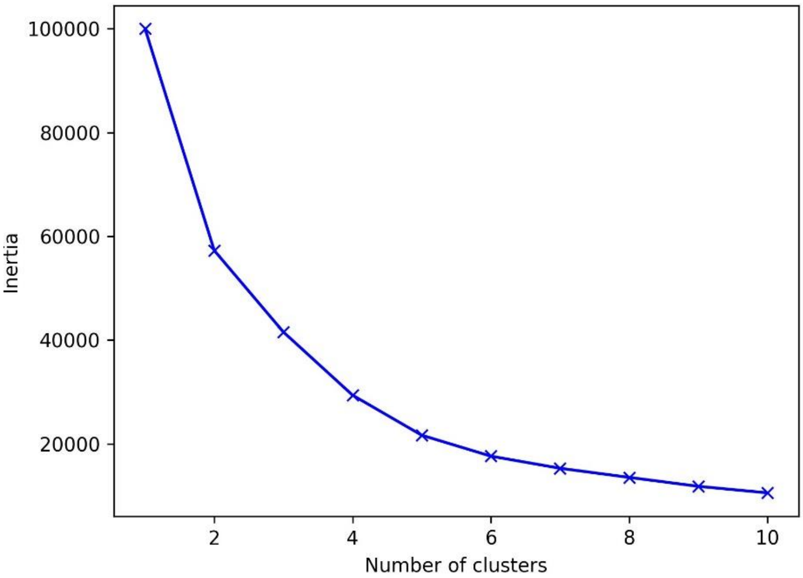


3 is the optimal number of clusters to use here as it is the point of the elbow.

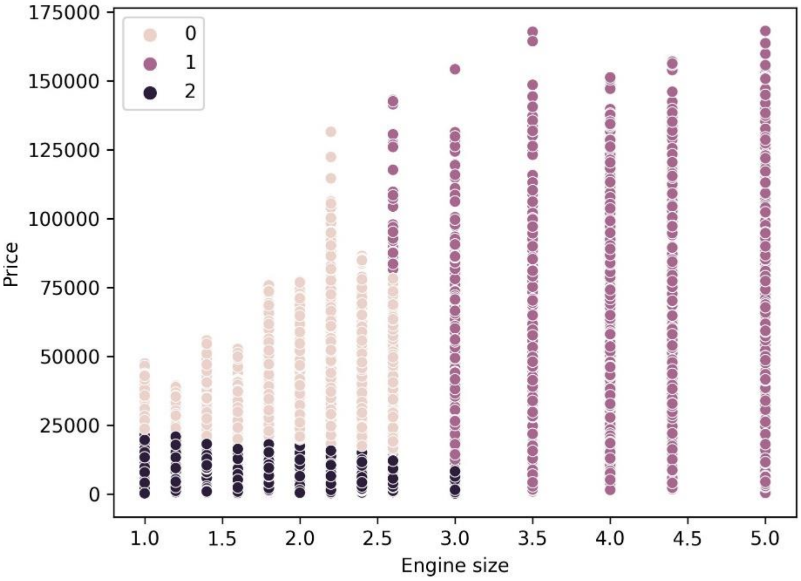


Davies Bouldin Index: 0.6866

Silhouette Coefficient: 0.5139
Price v Engine size



There is a confusion here between using 2 or 3 as the optimal number of clusters as both fall on the elbow and on computation none gave a lower DB index and higher Silhouette Coefficient. I will go for 3 to help for uniformity in comparing the model to be adopted against other variables. The difference between the DB index and the Silhouette Coefficient for selecting 2 or 3 is not much. It has reached the optimal number of clusters.

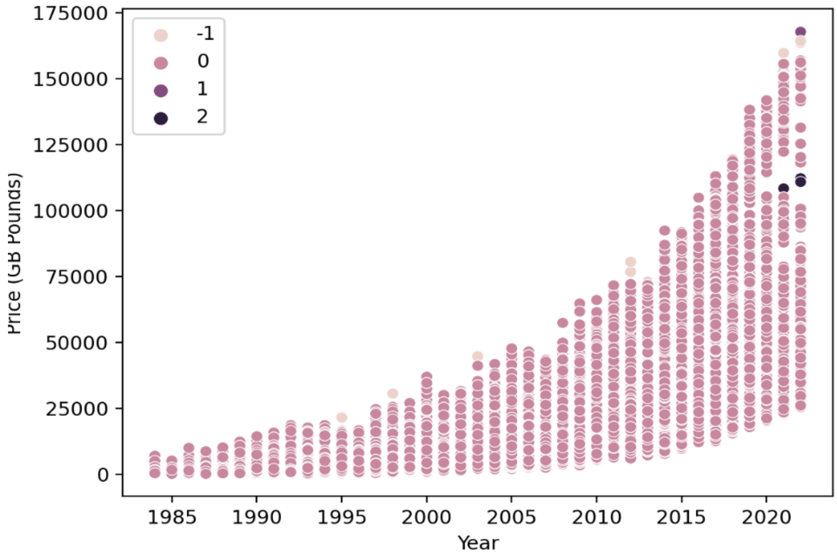


Davies Bouldin Index: 0.8982
Silhouette Coefficient: 0.4727

Answer 1f: The combination of Price and Year of Manufacture produces the best clustering result as it has a higher Silhouette Coefficient of 0 .51 and lesser DB index of 0.68 compared to combination of oth er variables as a model with Silhouette Coefficient nearer to 1 and DB index further from 1 is better (Theobald,2017).

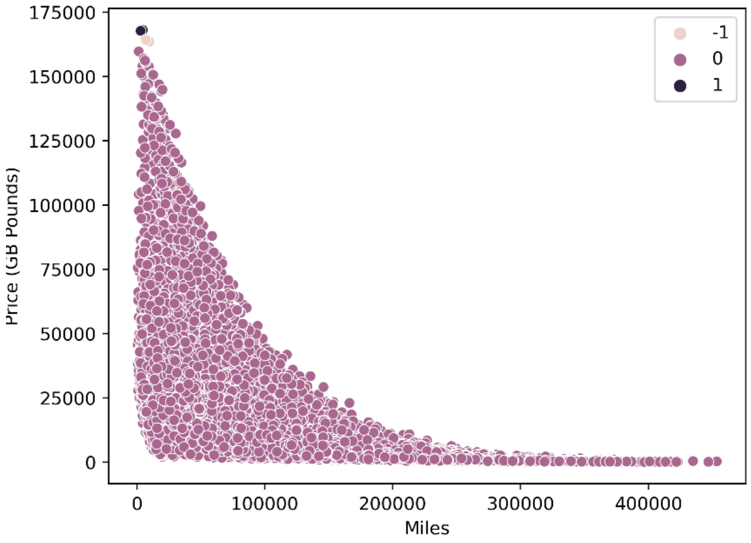
1g. DBSCAN

Price v Year of Manufacture



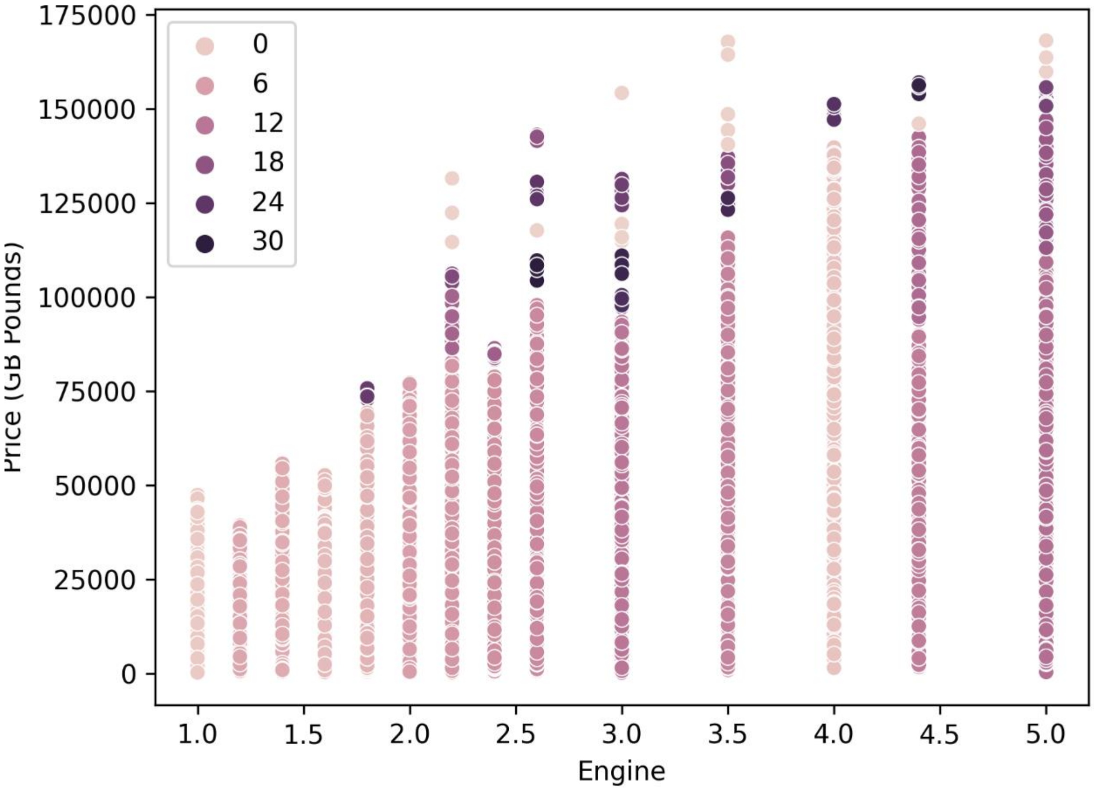
Davies Bouldin Index: 1.4952
Silhouette Coefficient: 0.6763

Price v Mileage



Davies Bouldin Index: 0.1570
Silhouette Coefficient: 0.8178

Price v Engine size



Davies Bouldin Index: 2.6414
Silhouette Coefficient: 0.0608

Answer 1fg: Comparing the results of the k-Means clustering model with result from DBSCAN clustering algorithm, DBSCAN produces the better result as seen in Price v Mileage with a low DB index of 0 .16 and a high of 0.82. The silhouette coefficient varies from −1 to +1. A value close to +1 means the point xi is much closer to the points in its own cluster. This indicates good clustering.

References

Igual, L., & Segui, S., (2017). Introduction to Data Science. Springer International Publications Switzerland

Theobald, O., (2017). Machine Learning for Absolute Beginners, 2nd ed, Amazon. Great Britain