

Analysis

1. There is no much difference in mean and median ,so maybe their will be no outlier
2. Mileage,Enginesize, horsepower, resaleprice having negative values
3. Owner count column having 133 null values
4. car_age standard deviation is 5 , so most of the car having age around 5,6,4 years.
5. no duplicated data
6. based on the Box plot of numerical data, except ownercount and car_age other column having outliers
7. correlation of car_age and resaleprice is -0.477, which means both having medium negative correlation, car_age increase price decrease
8. correlation of mileage and resaleprice = -0.21, which means both having slightly negative correlation, mileage increase price decrease
9. correlation of enginesize and resaleprice = 0.20, which means both having slightly positive correlation, enginesize increase price increase
10. correlation of horsepower and resaleprice = 0.65, which means both having strong positive correlation, horsepower increase price increase

Catagorical column

11. Car condition, brand making few changes in resale price column
12. fuel, trasmission not doing any changes in resale price column

RESULT

Prediction number	Major changes	Error metrics			
1	1. Owner count column null value replaced by Mean of that column. 2. Categorical column encoded [Fuel type, brand, transmission, carcondition] 3. standardization		MSE	RMSE	MAPE
		Train	34,624,079.09	5,884.22	0.10
		Test	33,217,715.91	5,763.48	0.09
2.	1. I used SelectKBest to select best 5 columns Columns are (['CarAge', 'Mileage', 'EngineSize', 'Horsepower', 'CarCondition'])		MSE	RMSE	MAPE
		Train	34,788,331.04	5,898.16	0.10
		Test	34,788,331.04	5,751.92	0.09

3.	1. I took only Horse power columns as train set , because it is having strong positive correlation with resaleprice	<table><tr><th></th><th>MSE</th><th>RMSE</th><th>MAPE</th><th>r square</th></tr><tr><td>Train</td><td>86,740,700.29</td><td>9,313.47</td><td>0.14</td><td>-.0.38</td></tr><tr><td>Test</td><td>77,096,509.06</td><td>8,780.46</td><td>0.14</td><td>-0.23</td></tr></table>					MSE	RMSE	MAPE	r square	Train	86,740,700.29	9,313.47	0.14	-.0.38	Test	77,096,509.06	8,780.46	0.14	-0.23
	MSE	RMSE	MAPE	r square																
Train	86,740,700.29	9,313.47	0.14	-.0.38																
Test	77,096,509.06	8,780.46	0.14	-0.23																
4.	1.Based on the sns.pairplot(data) , i can see some outliers in the following columns [Car_age, Mileage,Engine_size] 2.Remove those outliers in car_age,mileage,enginesize columns 3.Encode catagorical columns 4.standardization	<table><tr><th></th><th>MSE</th><th>RMSE</th><th>MAPE</th><th>r-square</th></tr><tr><td>Train</td><td>34,823,019.02</td><td>5,901.10</td><td>0.09</td><td>0.65</td></tr><tr><td>Test</td><td>32,098,033.11</td><td>5,665.51</td><td>0.09</td><td>0.70</td></tr></table>					MSE	RMSE	MAPE	r-square	Train	34,823,019.02	5,901.10	0.09	0.65	Test	32,098,033.11	5,665.51	0.09	0.70
	MSE	RMSE	MAPE	r-square																
Train	34,823,019.02	5,901.10	0.09	0.65																
Test	32,098,033.11	5,665.51	0.09	0.70																