In [8]:
```
import numpy as np
import pandas as pd
```

In [11]:
```
train = pd.read_csv("C:\\Users\\HP\\Desktop\\original projectsimpli\\machine learning\\project mercedese benz\\train.csv")
train.head()
```

Out[11]:

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	...	X375	X376	X377	X378	X379	X380	X382	X383	X384	X385
0	0	130.81	k	v	at	a	d	u	j	o	...	0	0	1	0	0	0	0	0	0	0
1	6	88.53	k	t	av	e	d	y	l	o	...	1	0	0	0	0	0	0	0	0	0
2	7	76.26	az	w	n	c	d	x	j	x	...	0	0	0	0	0	0	1	0	0	0
3	9	80.62	az	t	n	f	d	x	l	e	...	0	0	0	0	0	0	0	0	0	0
4	13	78.02	az	v	n	f	d	h	d	n	...	0	0	0	0	0	0	0	0	0	0

5 rows × 378 columns

In [12]:
```
test = pd.read_csv("C:\\Users\\HP\\Desktop\\original projectsimpli\\machine learning\\project mercedese benz\\test.csv")
test.head()
```

Out[12]:

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	...	X375	X376	X377	X378	X379	X380	X382	X383	X384	X385
0	1	az	v	n	f	d	t	a	w	0	...	0	0	0	1	0	0	0	0	0	0
1	2	t	b	ai	a	d	b	g	y	0	...	0	0	1	0	0	0	0	0	0	0
2	3	az	v	as	f	d	a	j	j	0	...	0	0	0	1	0	0	0	0	0	0
3	4	az	l	n	f	d	z	l	n	0	...	0	0	0	1	0	0	0	0	0	0
4	5	w	s	as	c	d	y	i	m	0	...	1	0	0	0	0	0	0	0	0	0

5 rows × 377 columns

In [14]:
```
train.isnull().sum()
```

Out[14]:
```
ID      0
y       0
X0      0
X1      0
X2      0
       ..
X380    0
X382    0
X383    0
X384    0
X385    0
Length: 378, dtype: int64
```

In [15]:
```
test.isnull().sum()
```

Out[15]:
```
ID      0
X0      0
X1      0
X2      0
X3      0
       ..
X380    0
X382    0
X383    0
X384    0
X385    0
Length: 377, dtype: int64
```

In [17]:
```
train.describe()
```

Out[17]:

	ID	y	X10	X11	X12	X13	X14	X15	X16	X17	...	X375	X376	X377	X378	X379	X380	X
count	4209.000000	4209.000000	4209.000000	4209.0	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	...	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.00
mean	4205.960798	100.669318	0.013305	0.0	0.075077	0.057971	0.428130	0.000475	0.002613	0.007603	...	0.318841	0.057258	0.314802	0.020670	0.009503	0.008078	0.00
std	2437.608688	12.679381	0.114590	0.0	0.263547	0.233716	0.494867	0.021796	0.051061	0.086872	...	0.466082	0.232363	0.464492	0.142294	0.097033	0.089524	0.08
min	0.000000	72.110000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	2095.000000	90.820000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
50%	4220.000000	99.150000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
75%	6314.000000	109.010000	0.000000	0.0	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00
max	8417.000000	265.320000	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00

8 rows × 370 columns

In [18]:
```
test.describe()
```

Out[18]:

	ID	X10	X11	X12	X13	X14	X15	X16	X17	X18	...	X375	X376	X377	X378	X379	X380	4
count	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	...	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4
mean	4211.039202	0.019007	0.000238	0.074364	0.061060	0.427893	0.000713	0.002613	0.008791	0.010216	...	0.325968	0.049656	0.311951	0.019244	0.011879	0.008078	
std	2423.078926	0.136565	0.015414	0.262394	0.239468	0.494832	0.026691	0.051061	0.093357	0.100570	...	0.468791	0.217258	0.463345	0.137399	0.108356	0.089524	
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	2115.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	4202.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	6310.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	...	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	
max	8416.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

8 rows × 369 columns

In [31]:
```
train_target = train["y"]
df_train = train.drop(["y","ID"],axis=1)
df_train.head(5)
```

Out[31]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	...	X375	X376	X377	X378	X379	X380	X382	X383	X384	X385
0	k	v	at	a	d	u	j	o	0	0	...	0	0	1	0	0	0	0	0	0	0
1	k	t	av	e	d	y	l	o	0	0	...	1	0	0	0	0	0	0	0	0	0
2	az	w	n	c	d	x	j	x	0	0	...	0	0	0	0	0	0	1	0	0	0
3	az	t	n	f	d	x	l	e	0	0	...	0	0	0	0	0	0	0	0	0	0
4	az	v	n	f	d	h	d	n	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 376 columns

In [33]:
```
df_train.var().sort_values().head(5)
```

Out[33]:
```
X330    0.0
X297    0.0
X268    0.0
X290    0.0
X235    0.0
dtype: float64
```

In [68]:
```
import seaborn as sns
from sklearn.feature_selection import VarianceThreshold
variance = VarianceThreshold(threshold=0)
from sklearn.preprocessing import LabelEncoder
label = LabelEncoder
import matplotlib.pyplot as plt
%matplotlib inline
```

In [34]:
```
train_without_zero_var = variance.fit_transform(df_train.iloc[:,9:])
train_without_zero_var
```

Out[34]:
```
array([[0, 1, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [1, 1, 0, ..., 0, 0, 0],
       [0, 0, 1, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [35]:
```
labeled = df_train.iloc[:,0:8]
labeled.head()
```

Out[35]:

	X0	X1	X2	X3	X4	X5	X6	X8
0	k	v	at	a	d	u	j	o
1	k	t	av	e	d	y	l	o
2	az	w	n	c	d	x	j	x
3	az	t	n	f	d	x	l	e
4	az	v	n	f	d	h	d	n

In [36]:
```
labeled.nunique()
```

Out[36]:
```
X0    47
X1    27
X2    44
X3     7
X4     4
X5    29
X6    12
X8    25
dtype: int64
```

In [38]:
```
labeled_data = labeled.apply(label().fit_transform)
labeled_data.head()
```

Out[38]:

	X0	X1	X2	X3	X4	X5	X6	X8
0	32	23	17	0	3	24	9	14
1	32	21	19	4	3	28	11	14
2	20	24	34	2	3	27	9	23
3	20	21	34	5	3	27	11	4
4	20	23	34	5	3	12	3	13

In [39]:
```
labeled_data.var()
```

Out[39]:
```
X0    188.741938
X1     72.777974
X2    118.808135
X3      3.027295
X4      0.005461
X5     68.076236
X6      8.508730
X8     49.531868
dtype: float64
```

In [40]:
```
train_zero_var = pd.DataFrame(train_without_zero_var)
train_zero_var.head()
```

Out[40]:

	0	1	2	3	4	5	6	7	8	9	...	345	346	347	348	349	350	351	352	353	354
0	0	1	0	0	0	0	1	0	0	1	...	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0	...	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 355 columns

In [42]:
```
train_data = pd.concat([labeled_data,train_zero_var],axis=1)
train_data.head()
```

Out[42]:

	X0	X1	X2	X3	X4	X5	X6	X8	0	1	...	345	346	347	348	349	350	351	352	353	354
0	32	23	17	0	3	24	9	14	0	1	...	0	0	1	0	0	0	0	0	0	0
1	32	21	19	4	3	28	11	14	0	0	...	1	0	0	0	0	0	0	0	0	0
2	20	24	34	2	3	27	9	23	0	0	...	0	0	0	0	0	0	1	0	0	0
3	20	21	34	5	3	27	11	4	0	0	...	0	0	0	0	0	0	0	0	0	0
4	20	23	34	5	3	12	3	13	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 363 columns

In [44]:
```
train_data.isnull().any()
```

Out[44]:
```
X0     False
X1     False
X2     False
X3     False
X4     False
```

```
        ...
350    False
351    False
352    False
353    False
354    False
Length: 363, dtype: bool
```

In [45]:

```
test = test.drop(['ID'],axis=1)
test.head()
```

Out[45]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	...	X375	X376	X377	X378	X379	X380	X382	X383	X384	X385
0	az	v	n	f	d	t	a	w	0	0	...	0	0	0	1	0	0	0	0	0	0
1	t	b	ai	a	d	b	g	y	0	0	...	0	0	1	0	0	0	0	0	0	0
2	az	v	as	f	d	a	j	j	0	0	...	0	0	0	1	0	0	0	0	0	0
3	az	l	n	f	d	z	l	n	0	0	...	0	0	0	1	0	0	0	0	0	0
4	w	s	as	c	d	y	i	m	0	0	...	1	0	0	0	0	0	0	0	0	0

5 rows × 376 columns

In [46]:

```
test.nunique()
```

Out[46]:

```
X0      49
X1      27
X2      45
X3       7
X4       4
        ..
X380     2
X382     2
X383     2
X384     2
X385     2
Length: 376, dtype: int64
```

In [47]:

```
test.isnull().any()
```

Out[47]:

```
X0      False
X1      False
X2      False
X3      False
X4      False
        ...
X380    False
X382    False
X383    False
X384    False
X385    False
Length: 376, dtype: bool
```

In [48]:

```
test.var().sort_values().head()
```

Out[48]:

```
X295    0.0
X369    0.0
X296    0.0
X257    0.0
X258    0.0
dtype: float64
```

In [49]:

```
test_without_zero_var = variance.fit_transform(test.iloc[:,9:])
test_without_zero_var
```

Out[49]:

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 1, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [50]:

```
test_data = pd.DataFrame(test_without_zero_var)
test_data.head()
```

Out[50]:

	0	1	2	3	4	5	6	7	8	9	...	352	353	354	355	356	357	358	359	360	361
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	...	0	0	1	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0

5 rows × 362 columns

In [51]:

```
labeled = test.iloc[:,0:8]
labeled.head()
```

Out[51]:

	X0	X1	X2	X3	X4	X5	X6	X8
0	az	v	n	f	d	t	a	w
1	t	b	ai	a	d	b	g	y
2	az	v	as	f	d	a	j	j
3	az	l	n	f	d	z	l	n
4	w	s	as	c	d	y	i	m

In [52]:

```
labeled_data = labeled.apply(label().fit_transform)
labeled_data.head()
```

Out[52]:

	X0	X1	X2	X3	X4	X5	X6	X8
0	21	23	34	5	3	26	0	22
1	42	3	8	0	3	9	6	24
2	21	23	17	5	3	0	9	9
3	21	13	34	5	3	31	11	13
4	45	20	17	2	3	30	8	12

In [58]:

```
test_data1 = pd.concat([labeled_data, test_data],axis=1)
test_data1.head(5)
```

Out[58]:

	X0	X1	X2	X3	X4	X5	X6	X8	0	1	...	352	353	354	355	356	357	358	359	360	361
0	21	23	34	5	3	26	0	22	0	0	...	0	0	0	1	0	0	0	0	0	0
1	42	3	8	0	3	9	6	24	0	0	...	0	0	1	0	0	0	0	0	0	0

2	21	23	17	5	3	0	9	9	0	0	...	0	0	0	1	0	0	0	0	0	0
3	21	13	34	5	3	31	11	13	0	0	...	0	0	0	1	0	0	0	0	0	0
4	45	20	17	2	3	30	8	12	0	0	...	1	0	0	0	0	0	0	0	0	0

5 rows × 370 columns

In [59]:
```
test_data1.isnull().any()
```

Out[59]:
```
X0     False
X1     False
X2     False
X3     False
X4     False
       ...
357    False
358    False
359    False
360    False
361    False
Length: 370, dtype: bool
```

In [60]:
```
from sklearn.model_selection import train_test_split
```

In [61]:
```
x_train,x_test,y_train,y_test = train_test_split(train_data,train_target,random_state=42,test_size=0.3)
```

In [62]:
```
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

Out[62]:
```
((2946, 363), (1263, 363), (2946,), (1263,))
```

In [64]:
```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
```

In [76]:
```
x_train = pca.fit_transform(x_train)
x_test = pca.transform(x_test)
```

In [2]:
```
from sklearn import svm
from sklearn.metrics import r2_score,mean_squared_error
from xgboost import XGBRegressor
xgbr = XGBRegressor(random_state=42)
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: