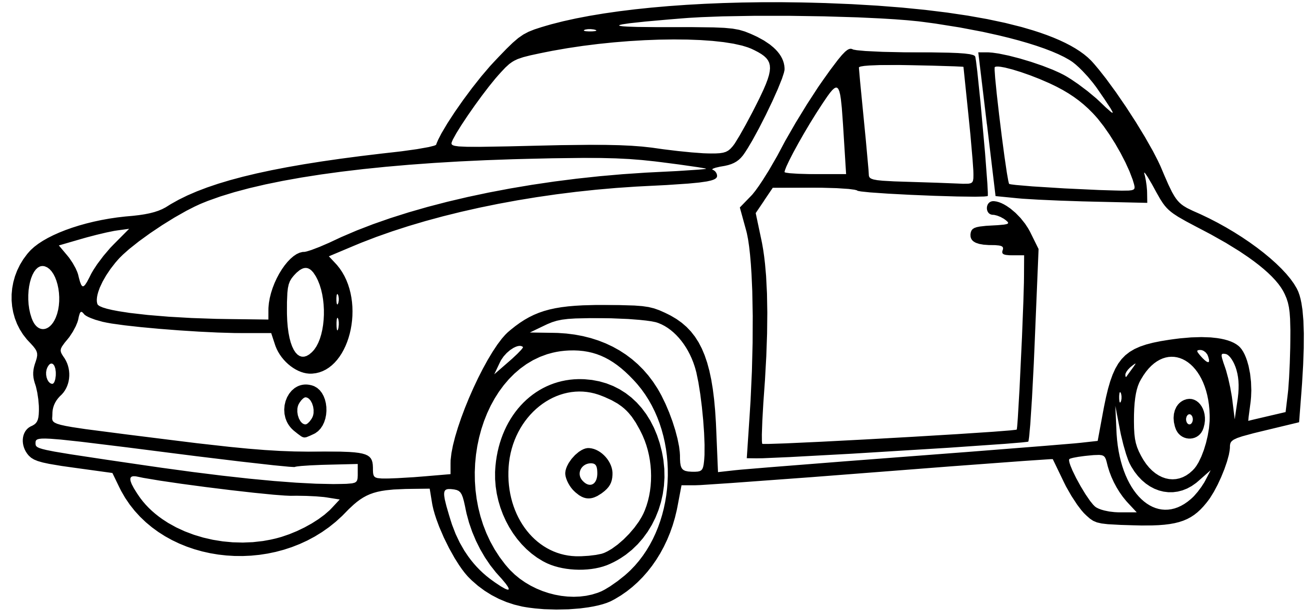
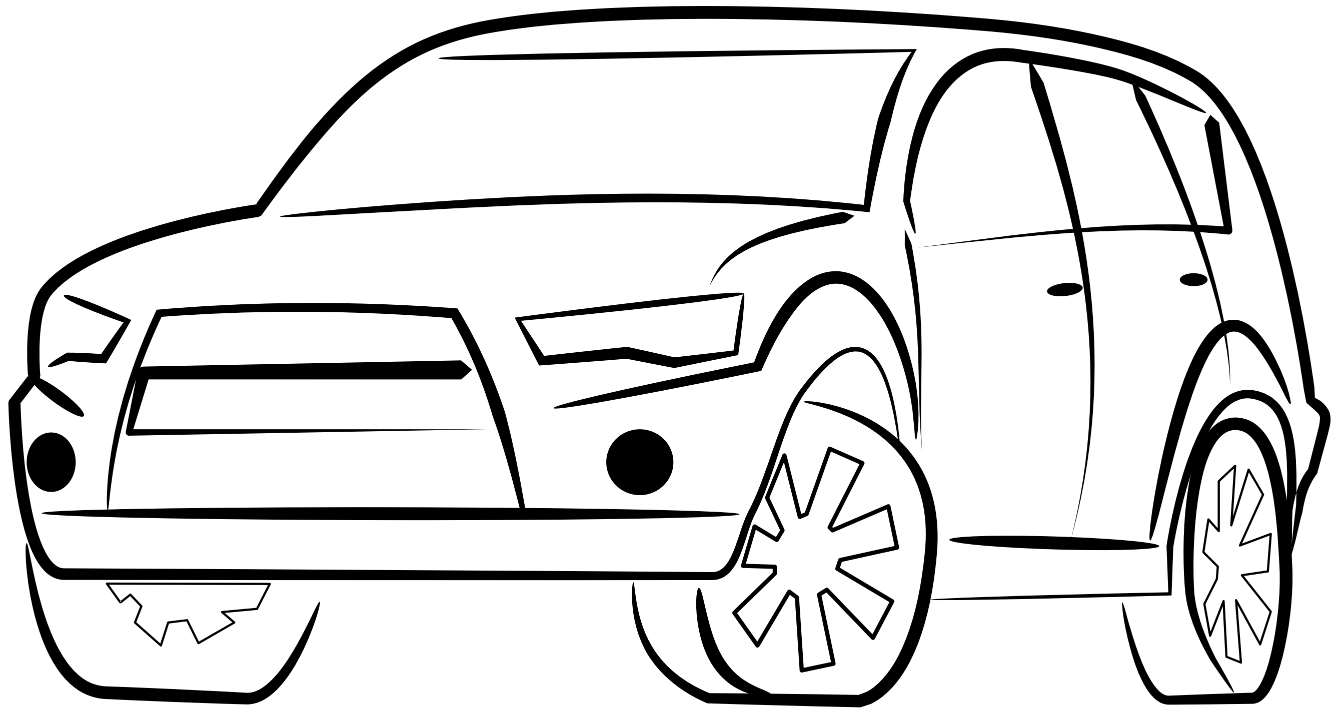
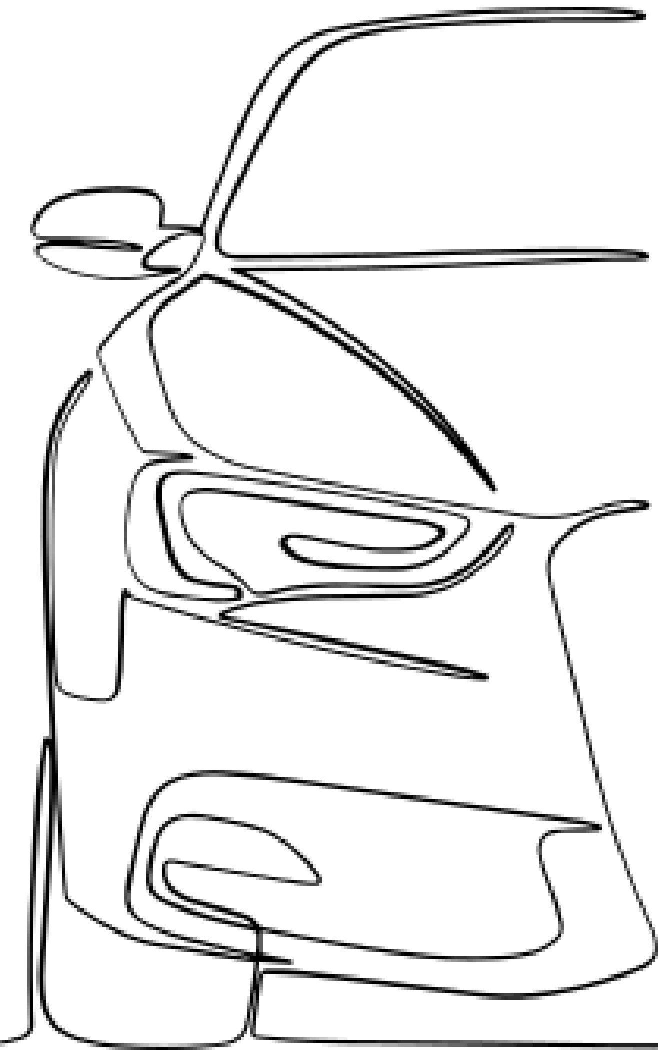


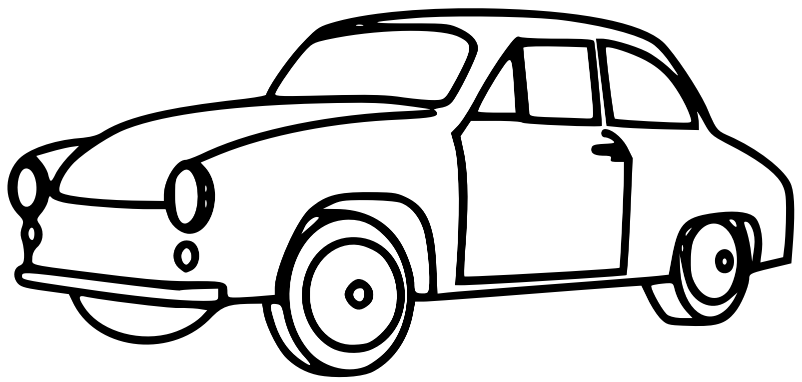
USED CAR PRICE PREDICTION



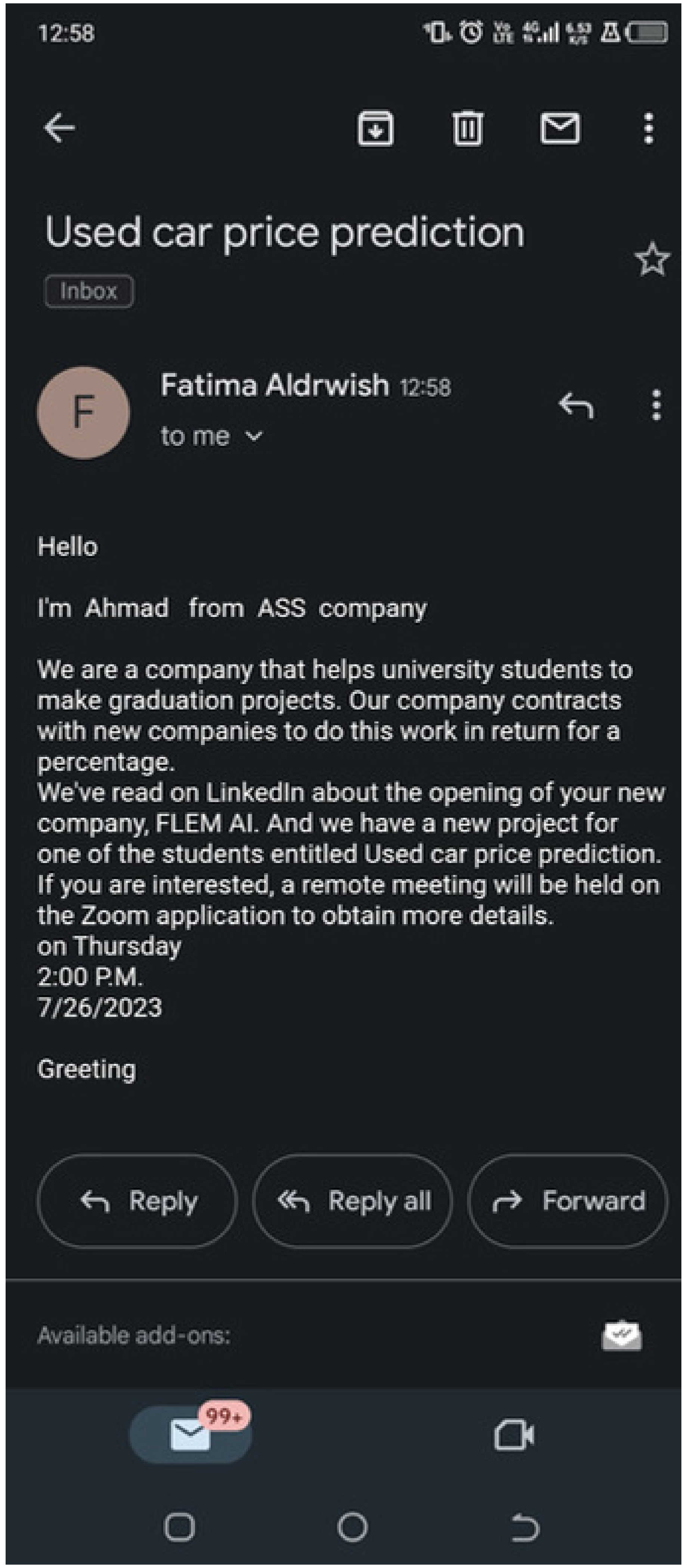


FLEM TEAM

- Fatima Aldrweesh
- LEEN ALMALLAH
- Eman Ababneh
- MARAM FAYEZ



BUSSINESS PROBLEM



12:58

Used car price prediction

Inbox

Fatima Aldrwish 12:58

to me

Hello

I'm Ahmad from ASS company

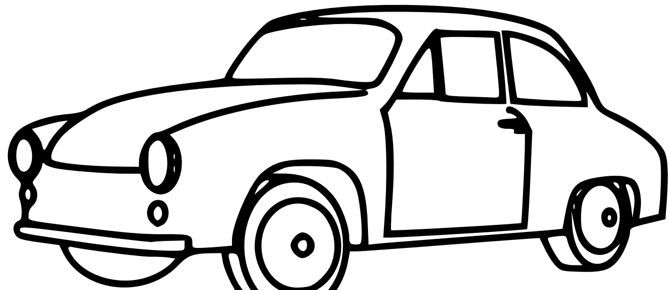
We are a company that helps university students to make graduation projects. Our company contracts with new companies to do this work in return for a percentage.
We've read on LinkedIn about the opening of your new company, FLEM AI. And we have a new project for one of the students entitled Used car price prediction.
If you are interested, a remote meeting will be held on the Zoom application to obtain more details.
on Thursday
2:00 P.M.
7/26/2023

Greeting

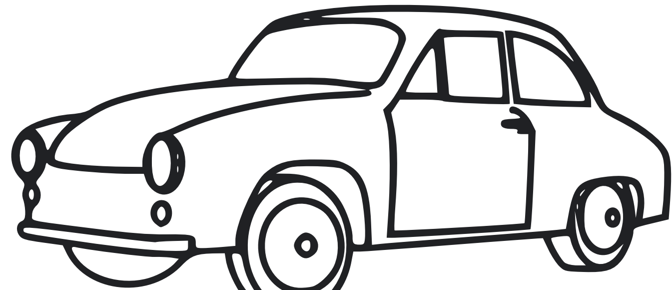
Reply Reply all Forward

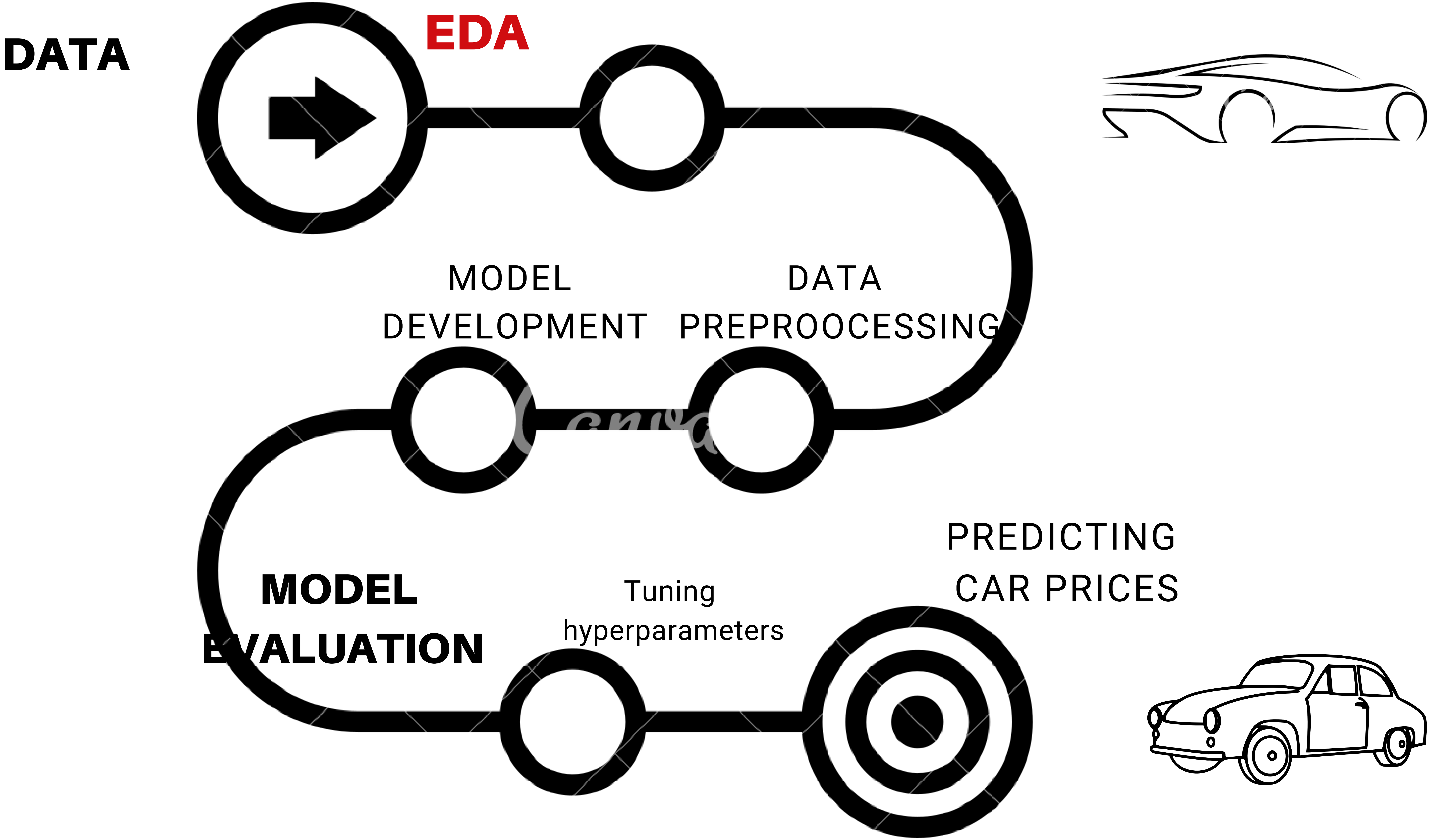
Available add-ons:

99+



Day	Schedule
Thursday (26/7/2023)	Meeting with ASS company
Friday	Team meeting (45 min) starting EDA (each one alone).
Saturday	Team meeting (45 min) Discussion about data, determine the work of each one.
Monday	Team meeting (45 min) Discussion about our work, determine the ML model to predict data.
Tuesday	Team meeting (45 min) Discussion about our work, determine the best models.
Wednesday	Team meeting (45 min) determine the procesure of presentation
Thursday	presentation




DATA
EDA
MODEL
DEVELOPMENT
DATA
PREPROOCESSING
MODEL
EVALUATION
Tuning
hyperparameters
PREDICTING
CAR PRICES

DATA

TRAIN

290129 * 20

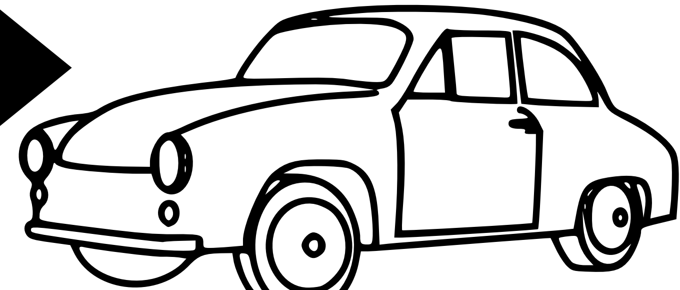
- float64(4)
- object(13)
- INT64(2)

TEST

124341 * 19

- float64(4)
- object(13)
- int64(1)

Zero duplicate value



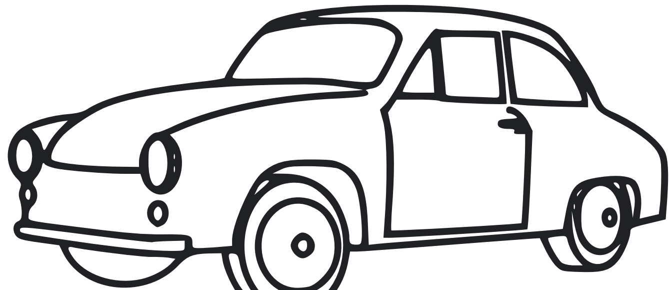
STATISTICAL DETAILS ABOUT NUMERIC COLUMNS

	UNNAMED: 0	ID	YEAR	ODOMETER	LAT	LONG	PRICE
COUNT	290129.0	2.901290e+05	290129.0	2.901290e+05	285726.0	285726.0	2.901290e+05
MEAN	207301.718	7.311503e+09	2011.359	9.764241e+04	38.505649	-94.61642	5.193300e+04
STD	119595.64	4.378450e+06	9.149422	2.058970e+05	5.830007	18.319158	9.591680e+06
MIN	0.0	7.301583e+09	1900	000e+0	-84.122245	-159.827728	0.0000e+00
25%	103622	7.308154e+09	2008	3.80e+04	34.60	-111.924900	5.9910e+03
50%	207440	7.312664e+09	2014	8.56150e+05	39.170	-88.212494	1.39900e+04
75%	310804	7.315255e+09	2017	1.334360e+05	42.4084	-80.83000	2.6500e+04
MAX	414469	7.317101e+09	2022	1.000e+07	82.252826	173.885502	3.736929e+09

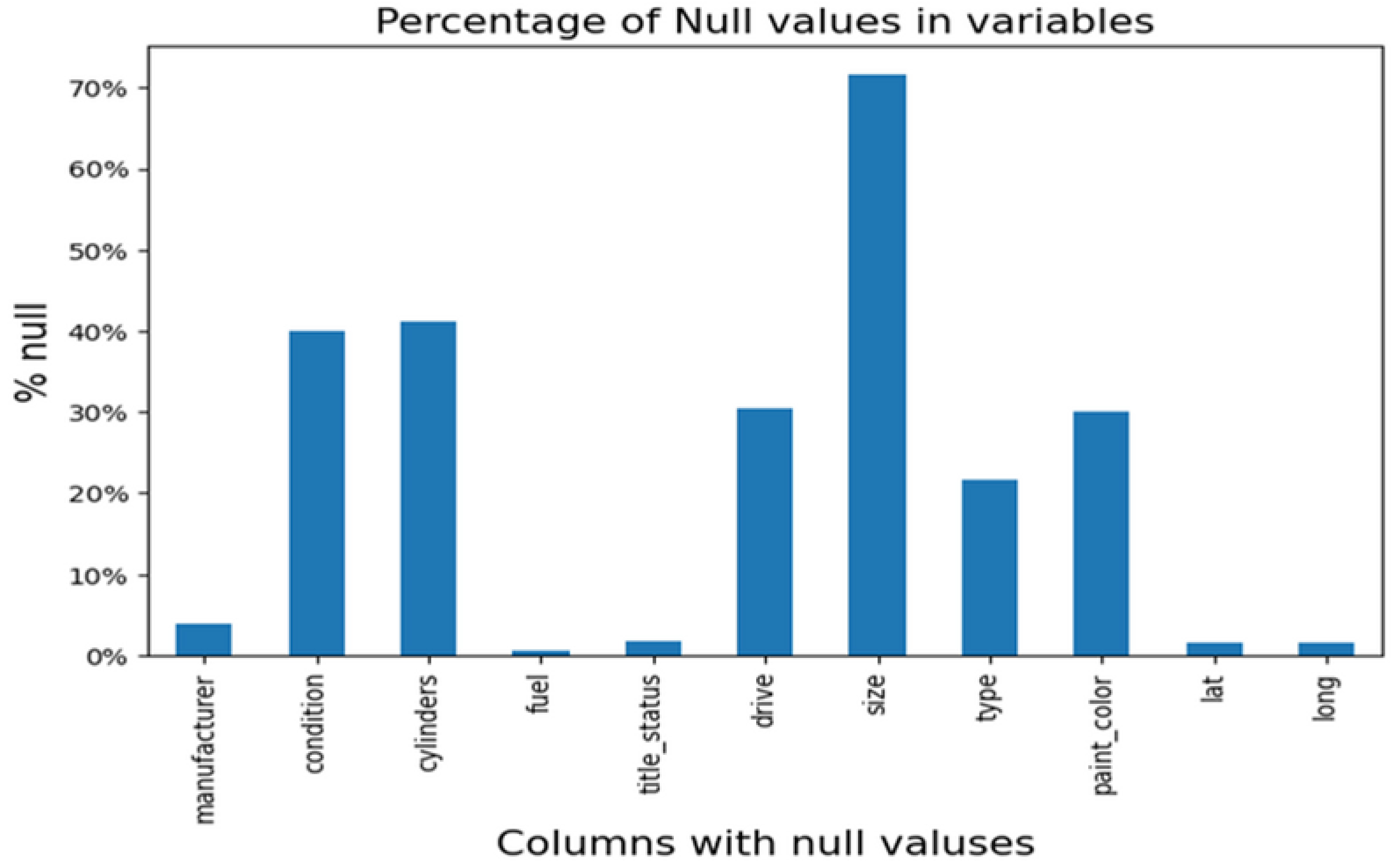
STATISTICAL DETAILS ABOUT CATEGORICAL COLUMNS

	MUNFACTURER	MODEL	CONDITION	CYLINDERS	FUEL	TITLE_STATUS	transmission
count	278527	289867	173807	170638	288414	284826	289867
uniqu	41	24300	6	8	5	6	3
top	ford	f-150	good	6 cylinders	gaf	clean	automatic
freq	48400	5493	84189	64759	242464	275588	229455

	DRIVE	TYPE	PAINT_COLOR	STATE	POSTING_DATE
count	201823	227283	202798	289867	289867
uniqu	3	13	12	51	267017
top	4wd	sedan	white	ca	2021-04-29T20:06:09
freq	89953	59452	54145	33963	8



THE PERCENTAGE OF NULL VALUES IN THE DATASET



Due to the prevalence of null values, exceeding 70% of the total column values, we will be removing the "size" column from our datasetrty!

DROP UNNECESSARY COLUMNS

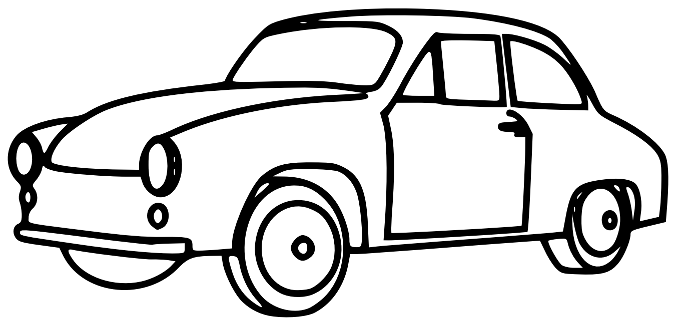
ID

LAT AND LONG

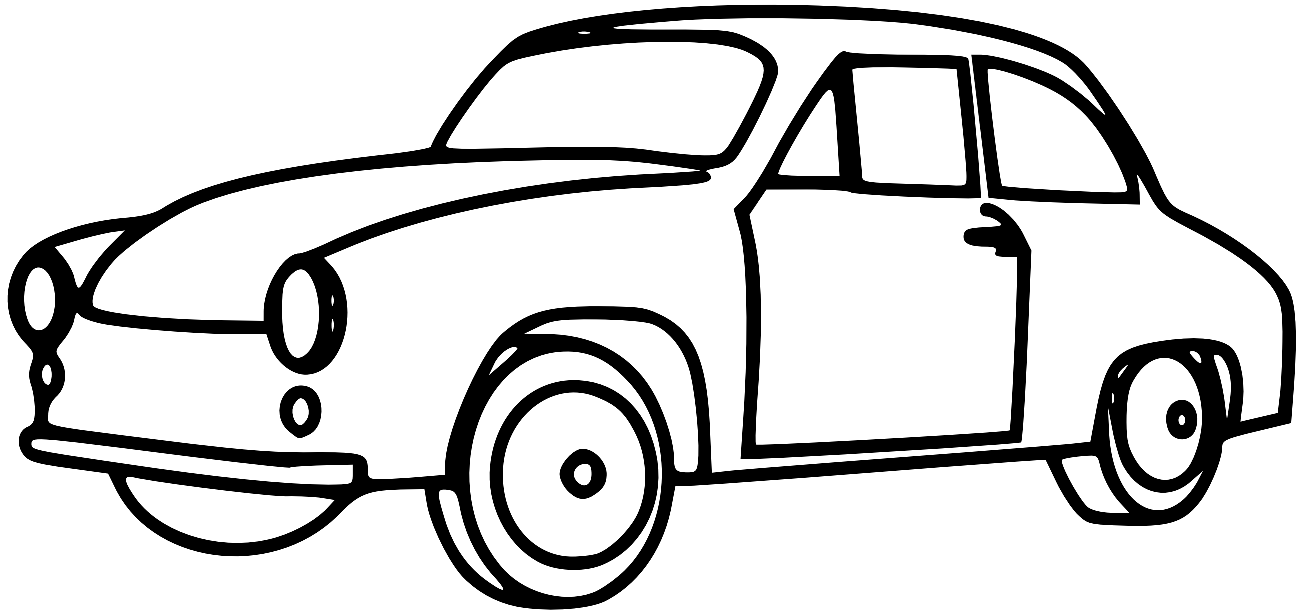
SIZE

Unnamed: 0

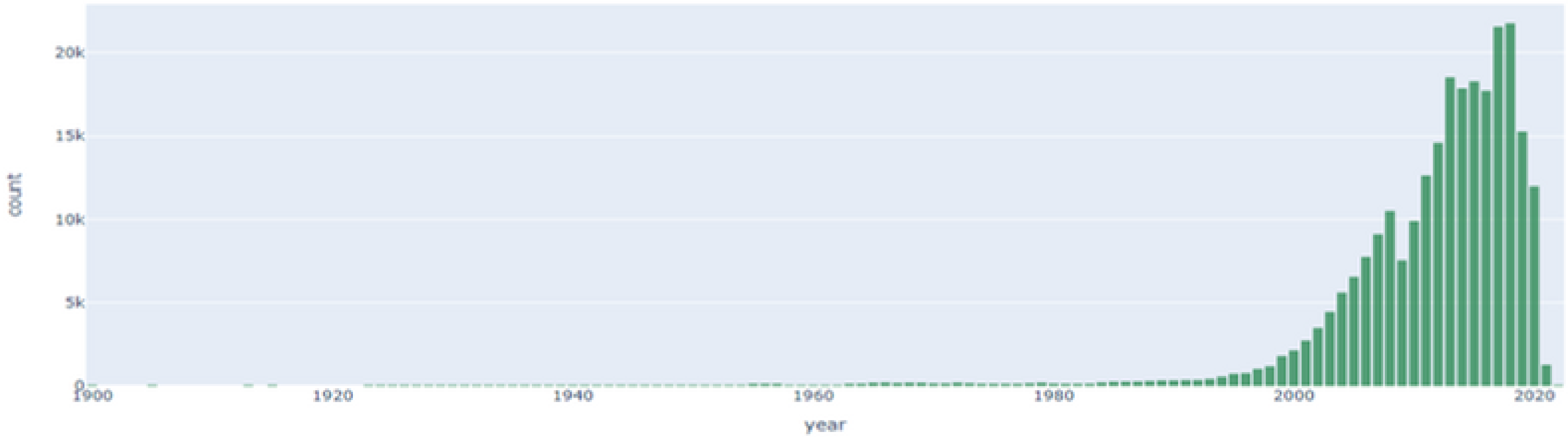
POSTING_DATE



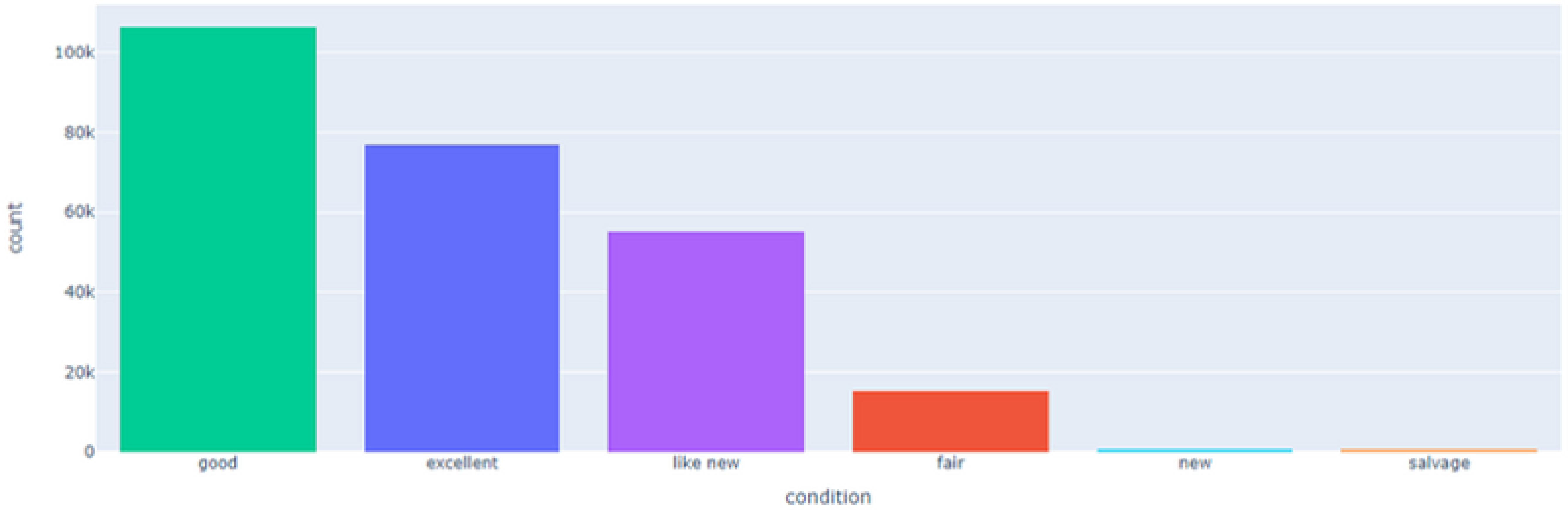
UUNIVARIATE ANALYSIS

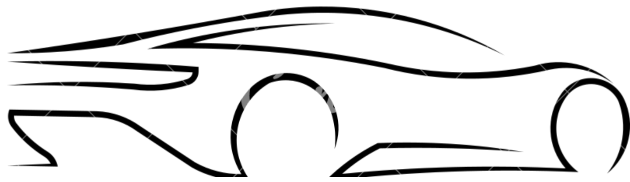


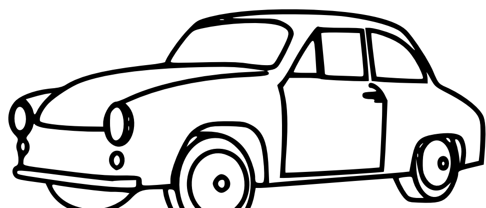
YEAR

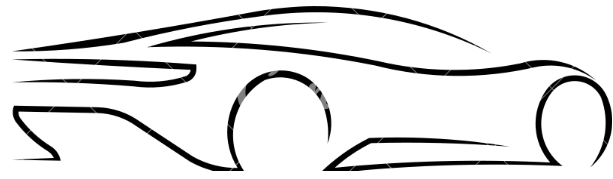


Condition

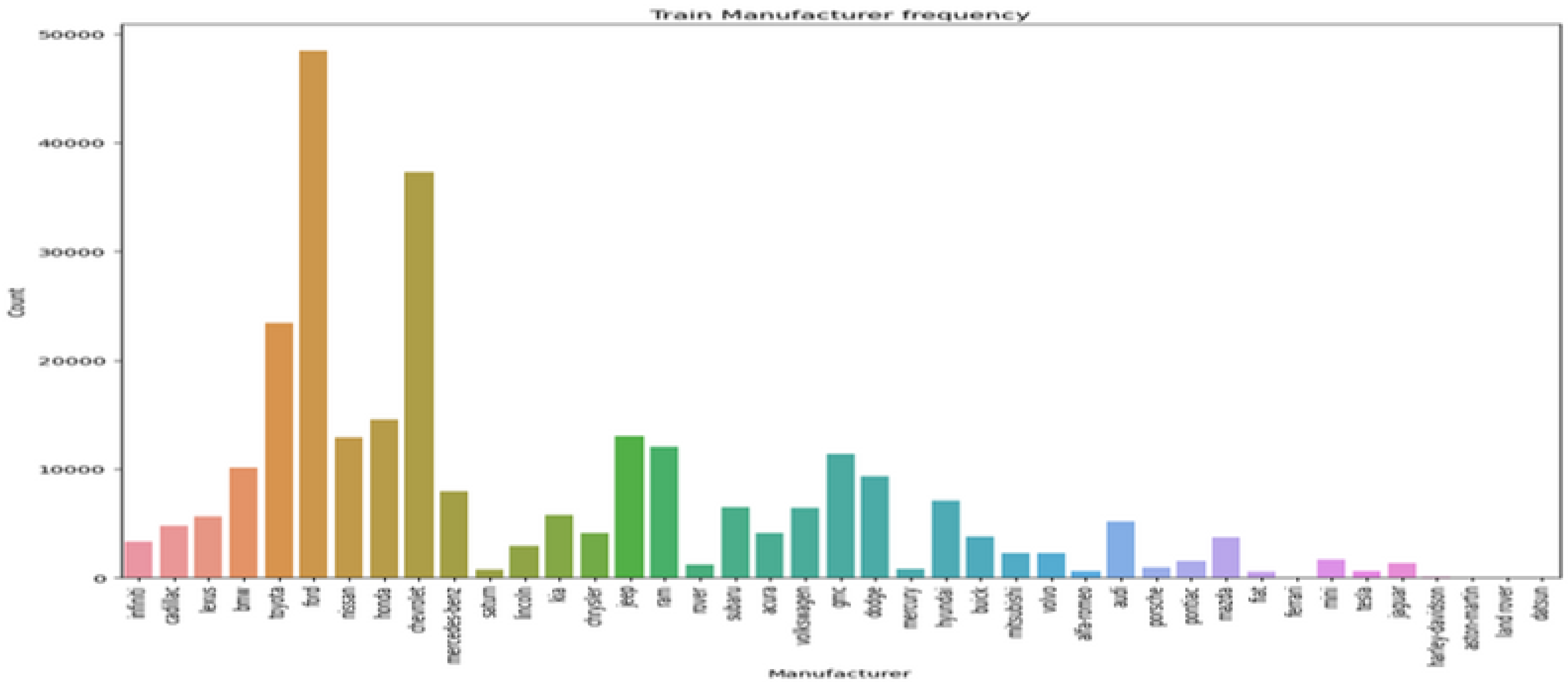




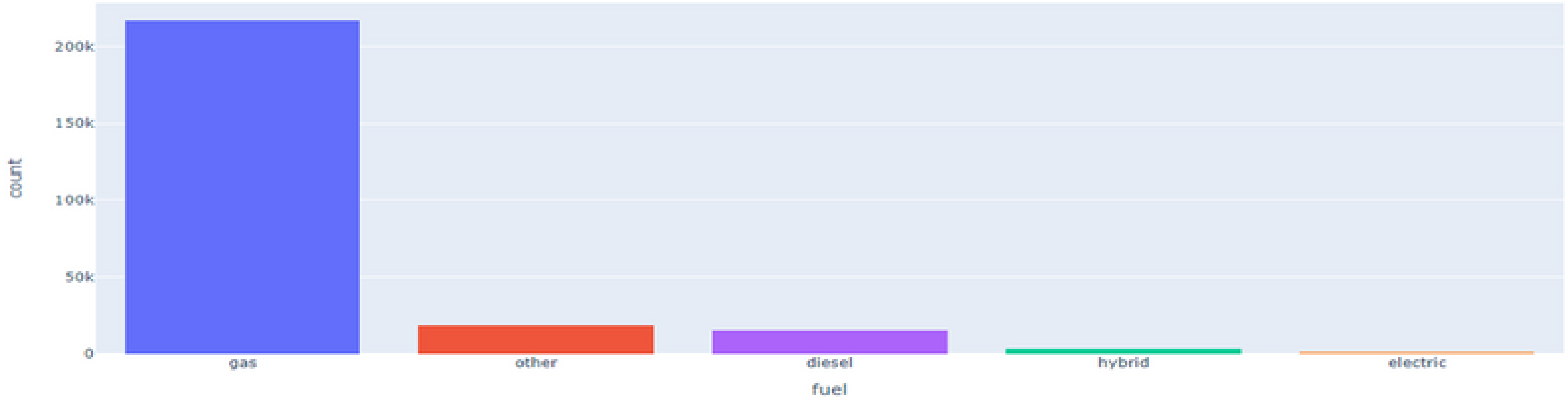


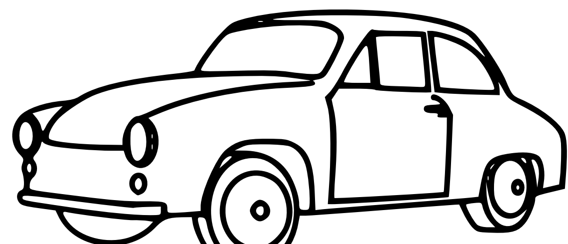


MANUFACTURER

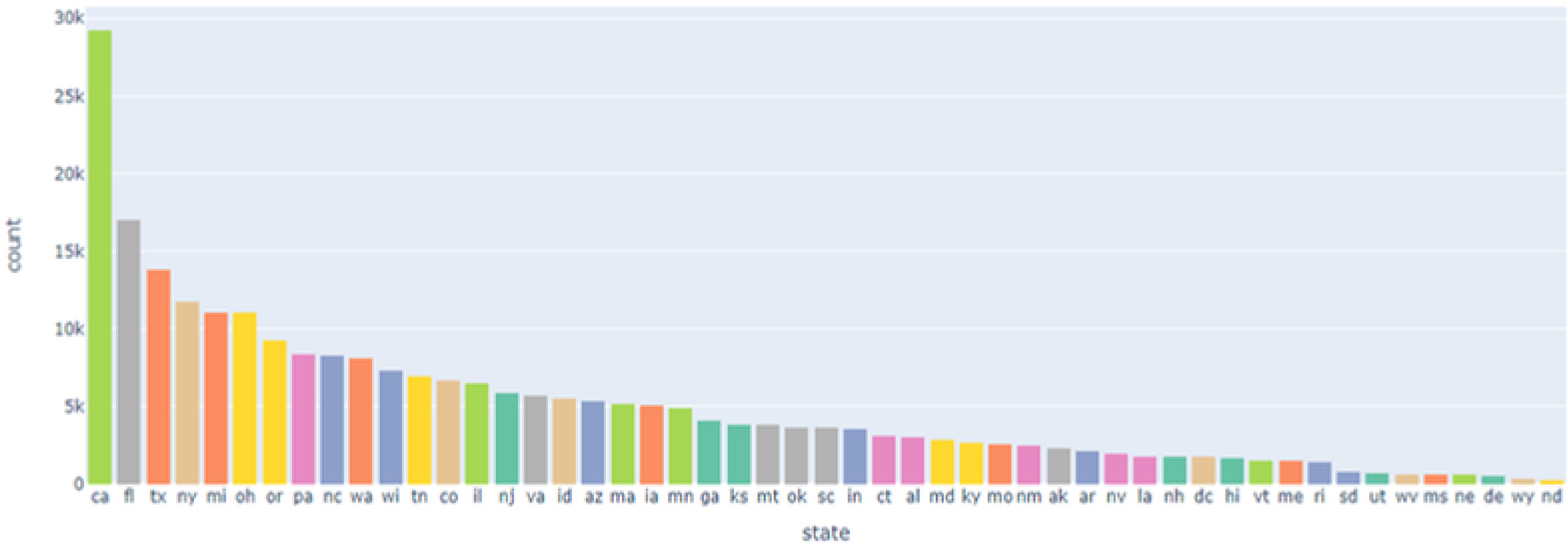


FUEL

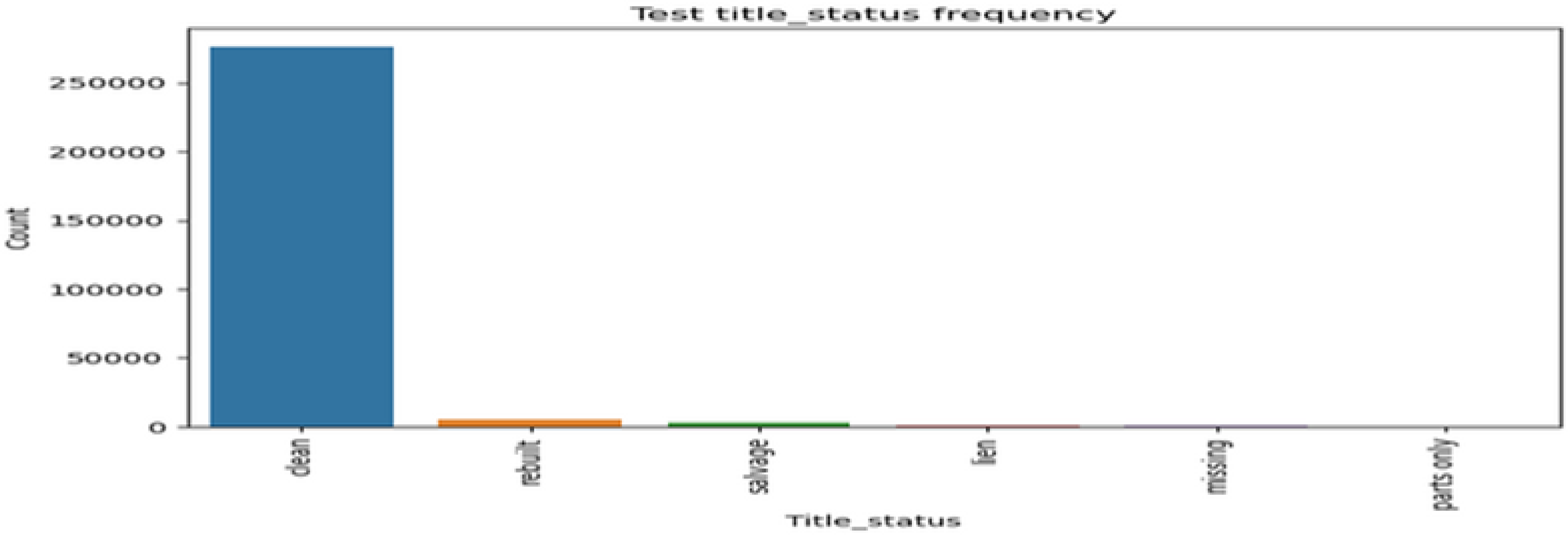




STATE

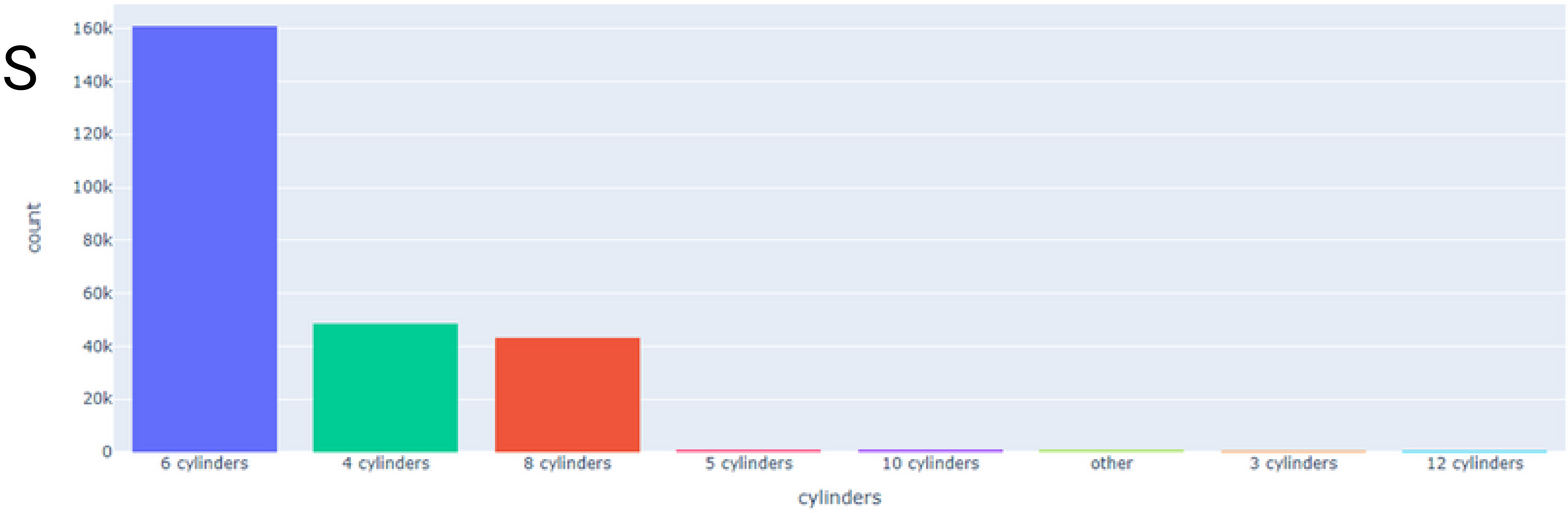


TITLE_STATUS

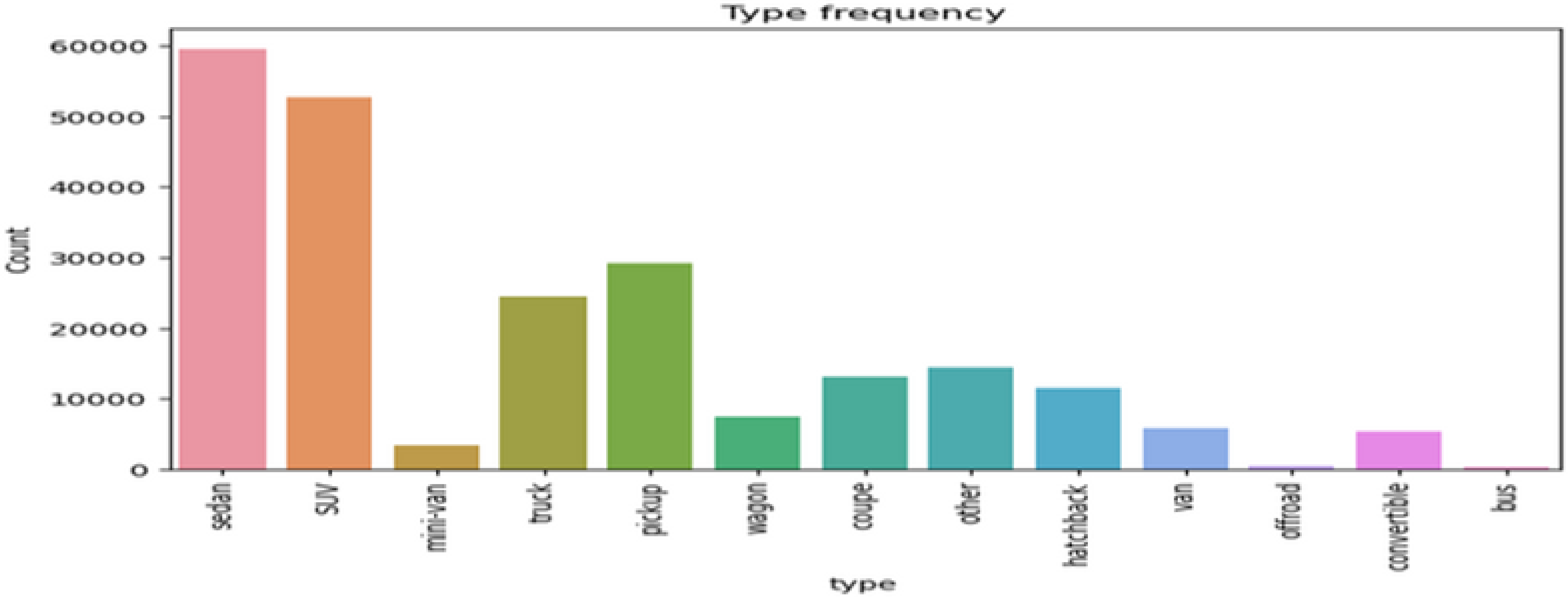


TRANSMISSION

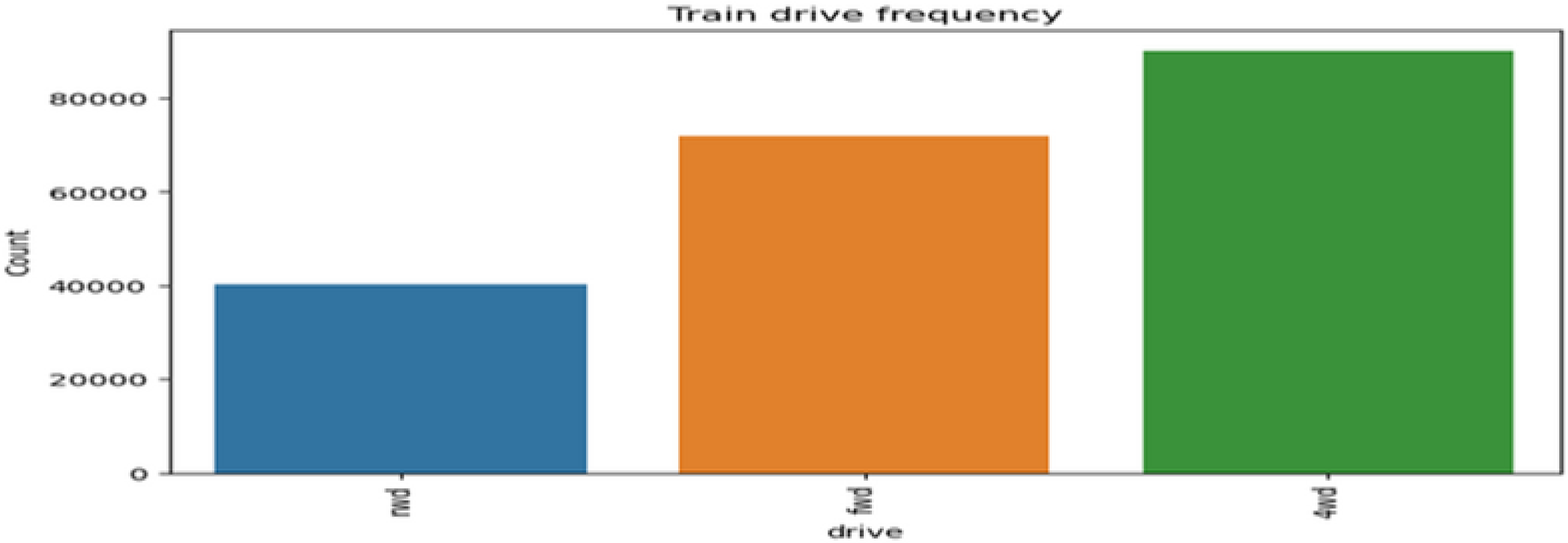
CYLINDERS



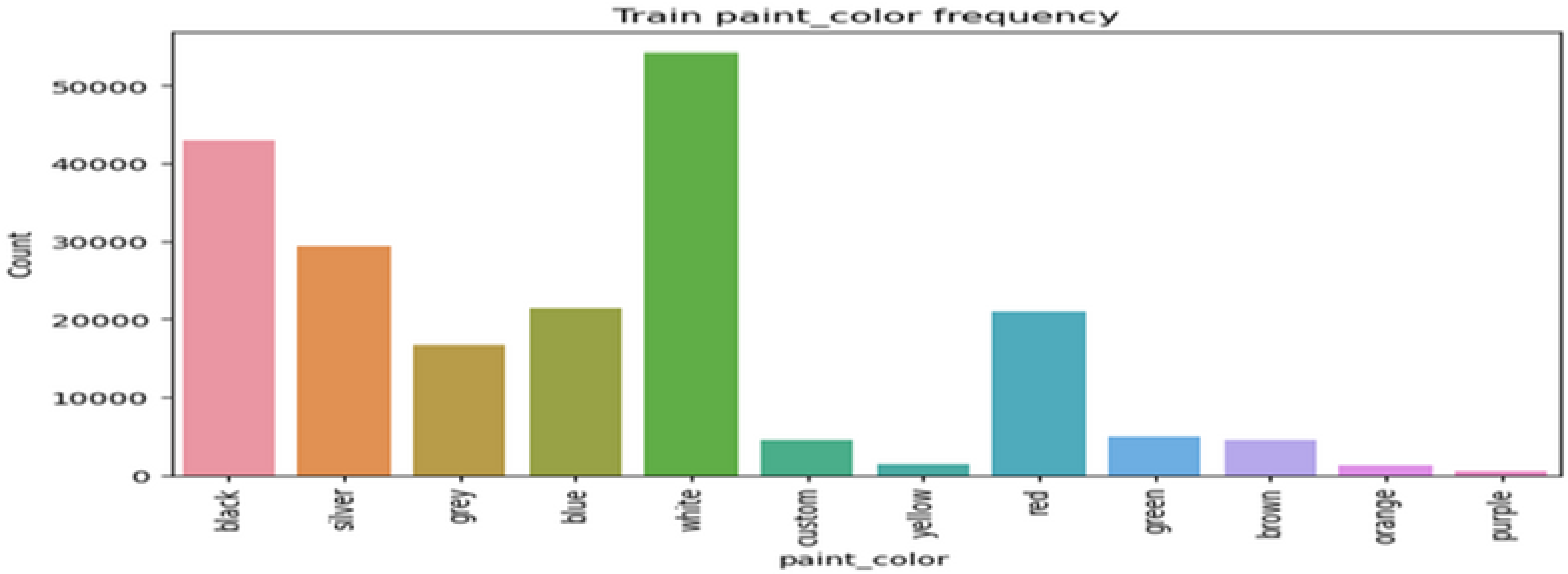
TYPE


Type frequency
60000
50000
40000
30000
20000
10000
0
Count
sedan
SUV
mini-van
truck
pickup
wagon
coupe
other
hatchback
van
offroad
convertible
bus
type

DRIVE

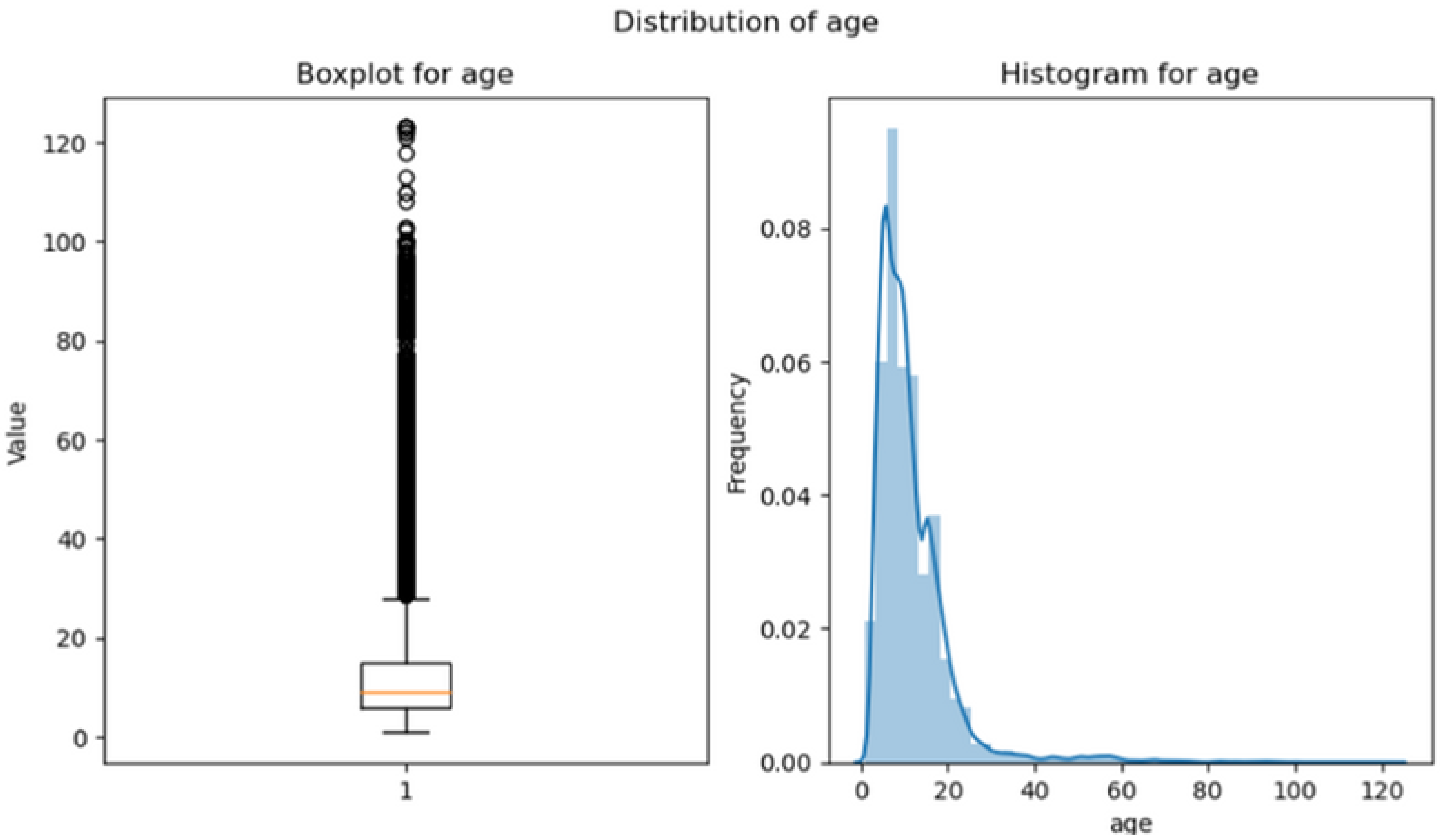

Train drive frequency
80000
60000
40000
20000
0
Count
rwd
fwd
4wd
drive

PAINT_COLOR



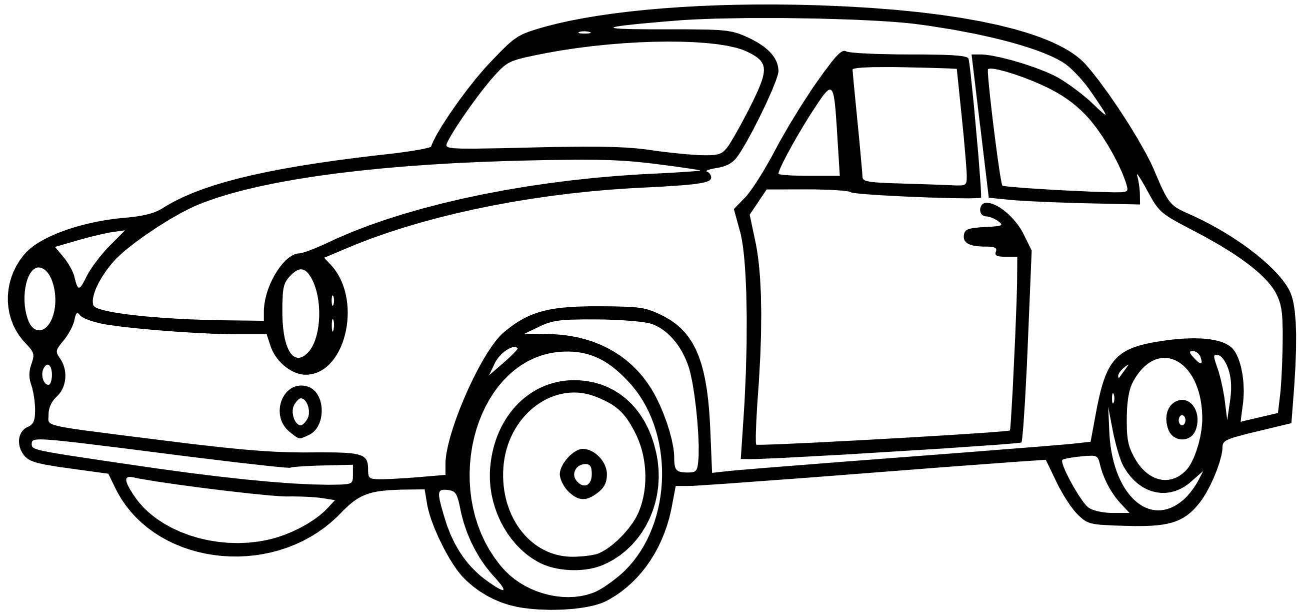
Train paint_color frequency
Count
50000
40000
30000
20000
10000
0
black
silver
grey
blue
white
custom
yellow
red
green
brown
orange
purple
paint_color

DATA WRANGLING AND CLEANING

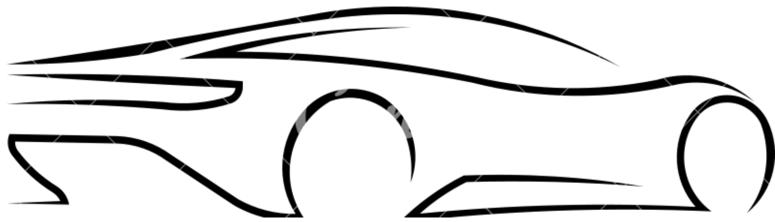


In order to better analyze and interpret our data, we will be generating a new column called "age" based on the information in the "year" column, and discarding the latter.

MISSING VALUE



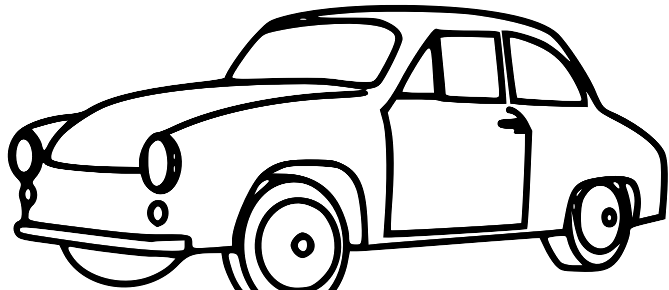
MANUFACTURER



We intend to utilize a model feature to manually fill in the null values in the "manufacturer" column, followed by filling the remaining null values with the column's mode, before finally dropping the "model" column.

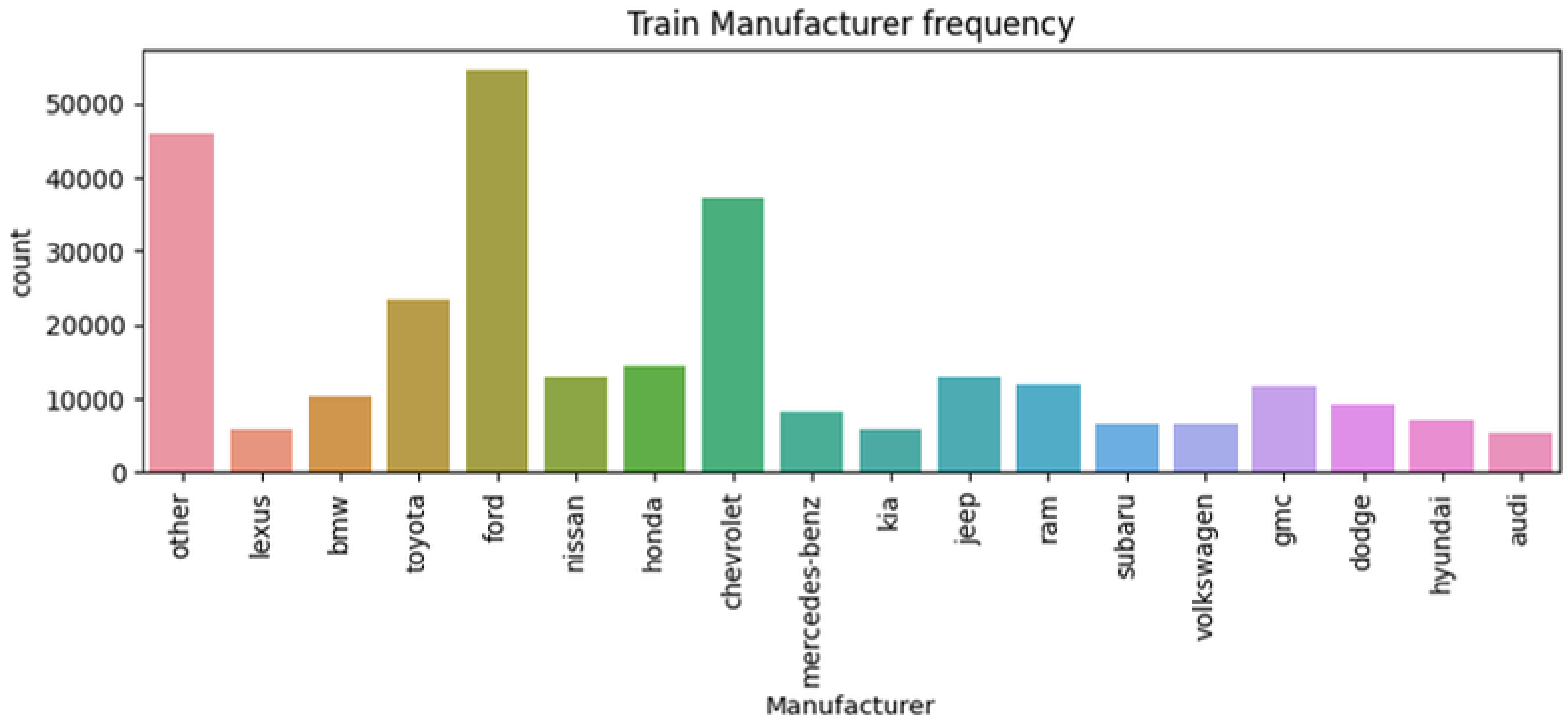
```
mask = (train.manufacturer.isnull()) & (train.model == 'Silverado k2500hd')
train.loc[mask, 'manufacturer'] = 'chevrolet'
mask = (test.manufacturer.isnull()) & (test.model == 'Silverado k2500hd')
test.loc[mask, 'manufacturer'] = 'chevrolet'

mask = (train.manufacturer.isnull()) & (train.model == 'Scion XB')
train.loc[mask, 'manufacturer'] = 'scion'
mask = (test.manufacturer.isnull()) & (test.model == 'Scion XB')
test.loc[mask, 'manufacturer'] = 'Scion'
```

```
mask = (train.manufacturer.isnull()) & (train.model == 'INTERNATIONAL WATER TRUCK')
train.loc[mask, 'manufacturer'] = 'navistar'
mask = (test.manufacturer.isnull()) & (test.model == 'INTERNATIONAL WATER TRUCK')
test.loc[mask, 'manufacturer'] = 'navistar'
```



MANUFACTURER

We intend to tidy our dataset by reclassifying "manufacturer" values with less than 5000 counts as "other", in addition to replacing "rover" with "land rover" and "general motors" with "gmc".

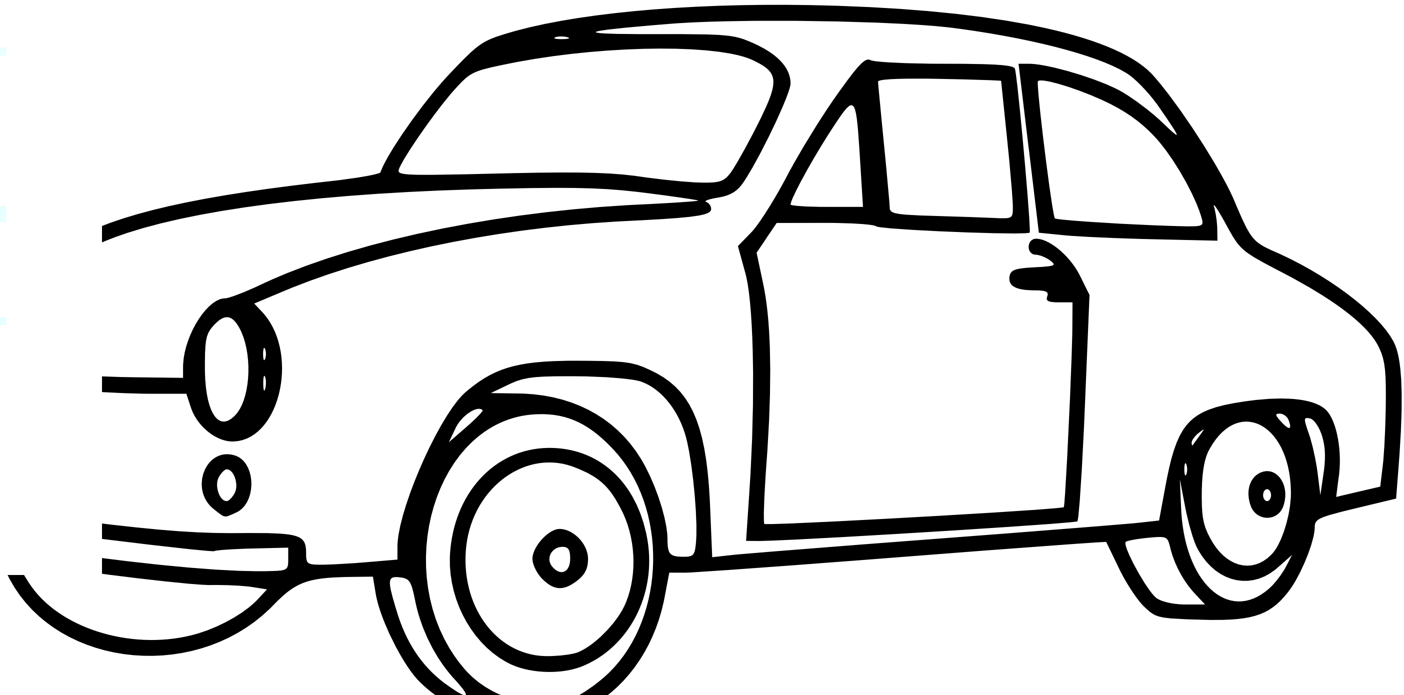


FUEL

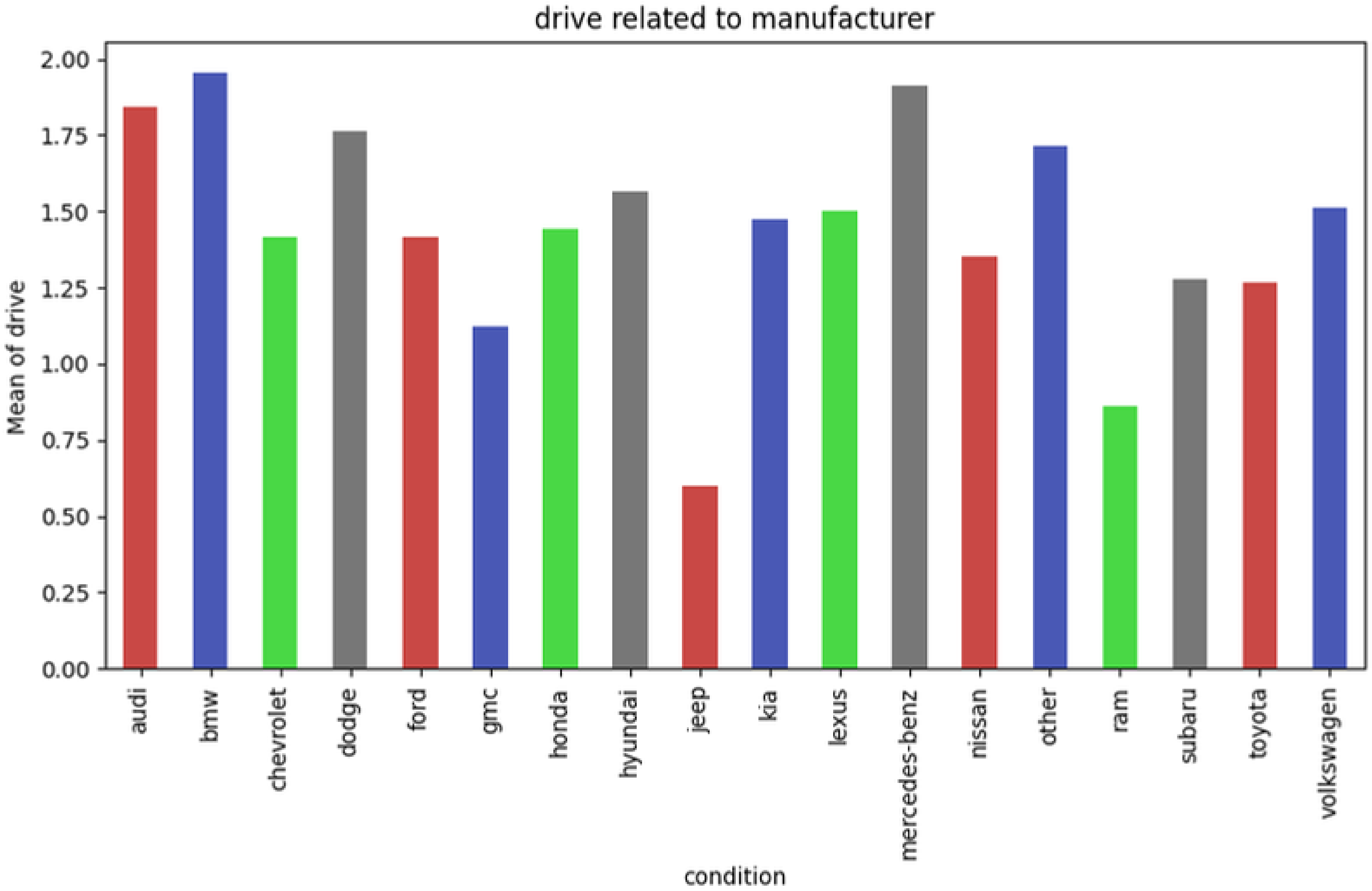
The "fuel" feature has (1453, 645) null values in the training and test datasets. To address this missing data, we will be filling these null values with an "other" value.

fuel	
gas	242693
other	20904
diesel	20309
hybrid	3607
electric	1163

fuel	
gas	242693
other	22357
diesel	20309
hybrid	3607
electric	1163

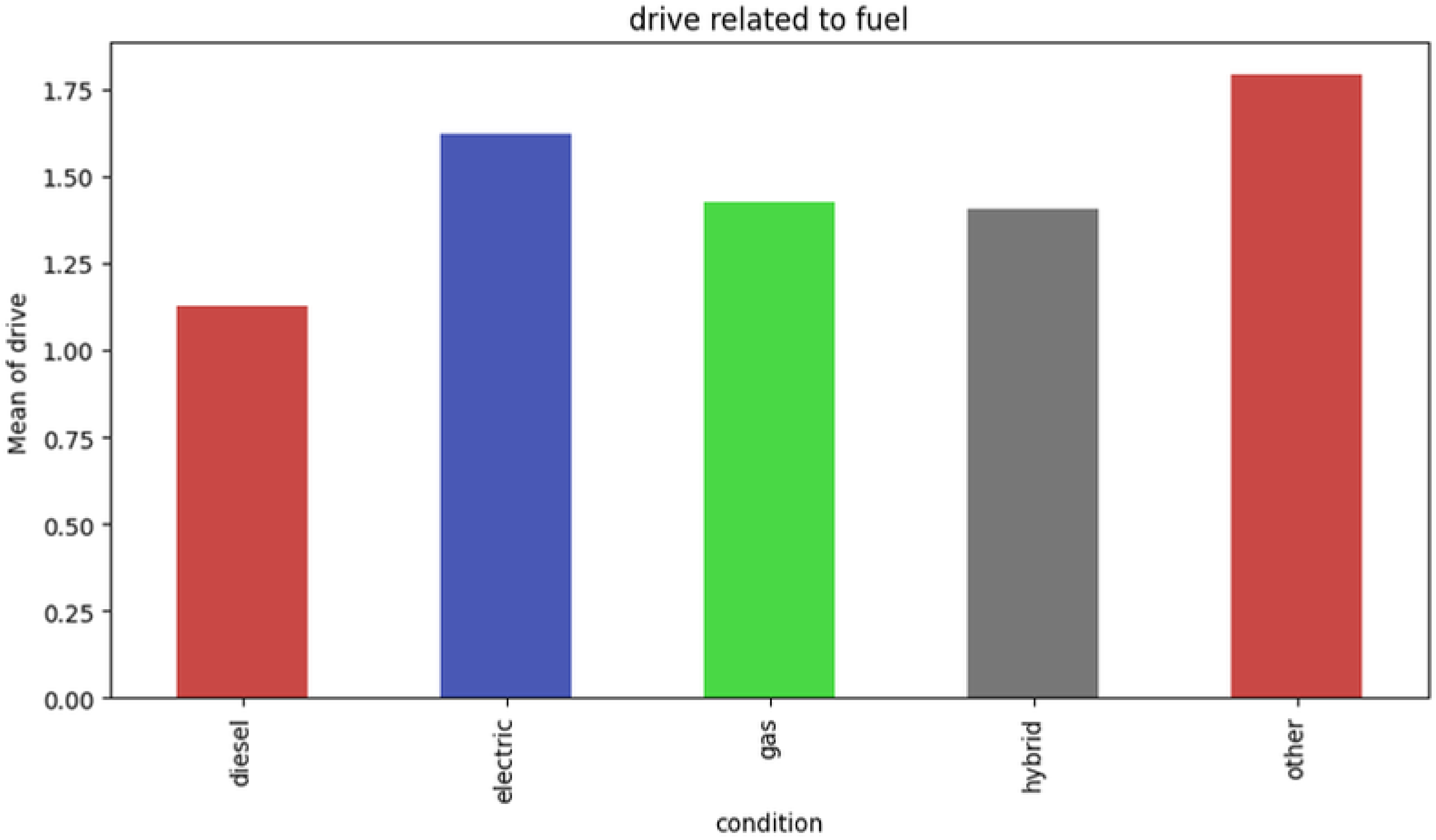


DRIVE


drive related to manufacturer
Mean of drive
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
audi
bmw
chevrolet
dodge
ford
gmc
honda
hyundai
jeep
kia
lexus
mercedes-benz
nissan
other
ram
subaru
toyota
volkswagen
condition

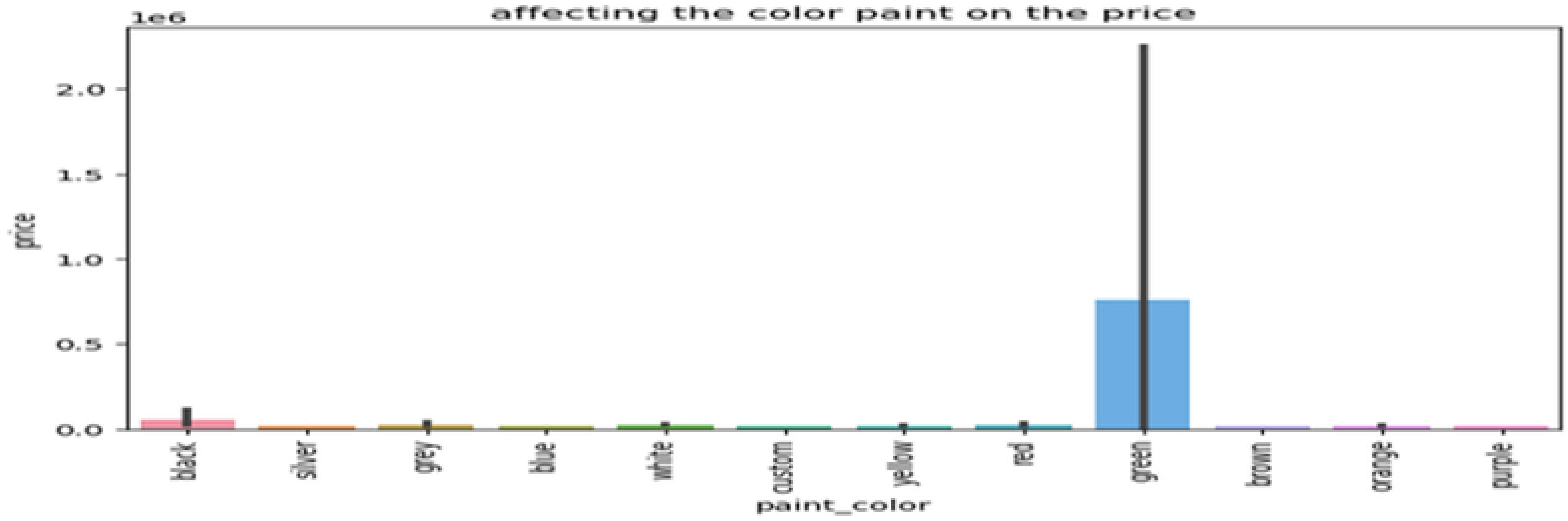
DRIVE

Based on previous figures:
'manufacturer‘, 'fuel‘
features affect on
‘drive’ feature so we use
them to fill the null values by
KNN algorithm,
we use search grid to find
the best value for K and we
obtain
(Best Parameters:
{'n_neighbors': 50}



PAINT COLOR

Based on this result the effect of color on price is few and there are (5004) value green is a little then we decided to drop this column.



CONDITION

THE "CONDITION" FEATURE OF A CAR IS INFLUENCED BY BOTH ITS "AGE" AND "ODOMETER" READINGS

By applying the KNN algorithm to fill in missing values in the "condition" feature and conducting a search grid to identify the most suitable value for K, we determined that the best parameters were {'n_neighbors': 1}

	condition	odometer
5	salvage	242048.583127
1	fair	212906.567069
0	excellent	105848.453411
3	like new	92809.418635
2	good	83121.556378
4	new	43749.556044

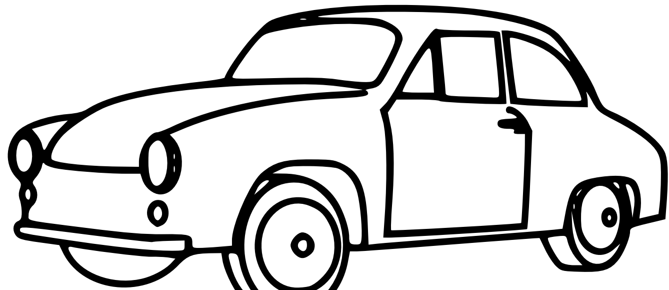
	condition	age
1	fair	24.941383
5	salvage	22.136476
0	excellent	12.940834
3	like new	11.344604
2	good	11.004555
4	new	9.352747

TITLE_STATUS

IN ‘TITLE_STATUS’ FEATURE WE FILL THE NULL VALUE BY MISSING VALUE

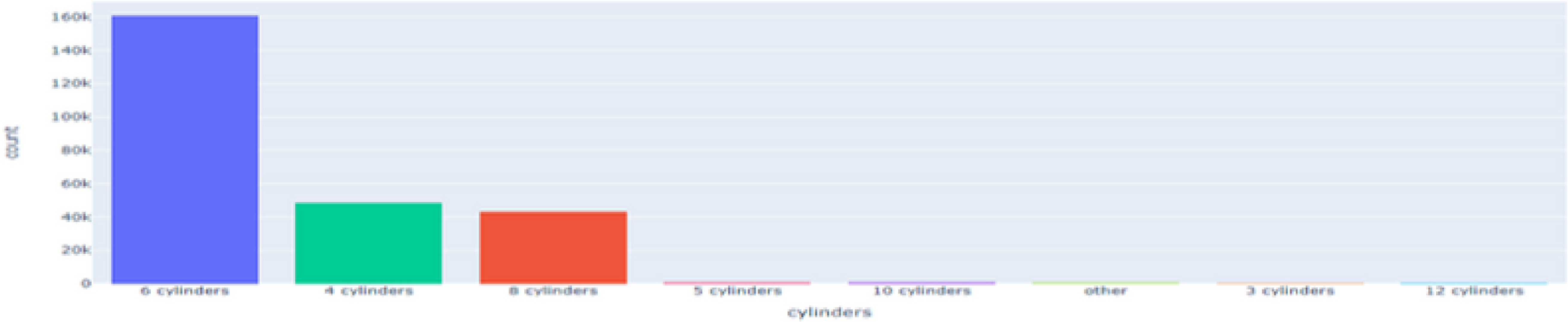
```
title_status
clean           275824
rebuilt           4924
salvage           2692
lien               981
missing            506
parts only         136
Name: count, dtype: int64
```

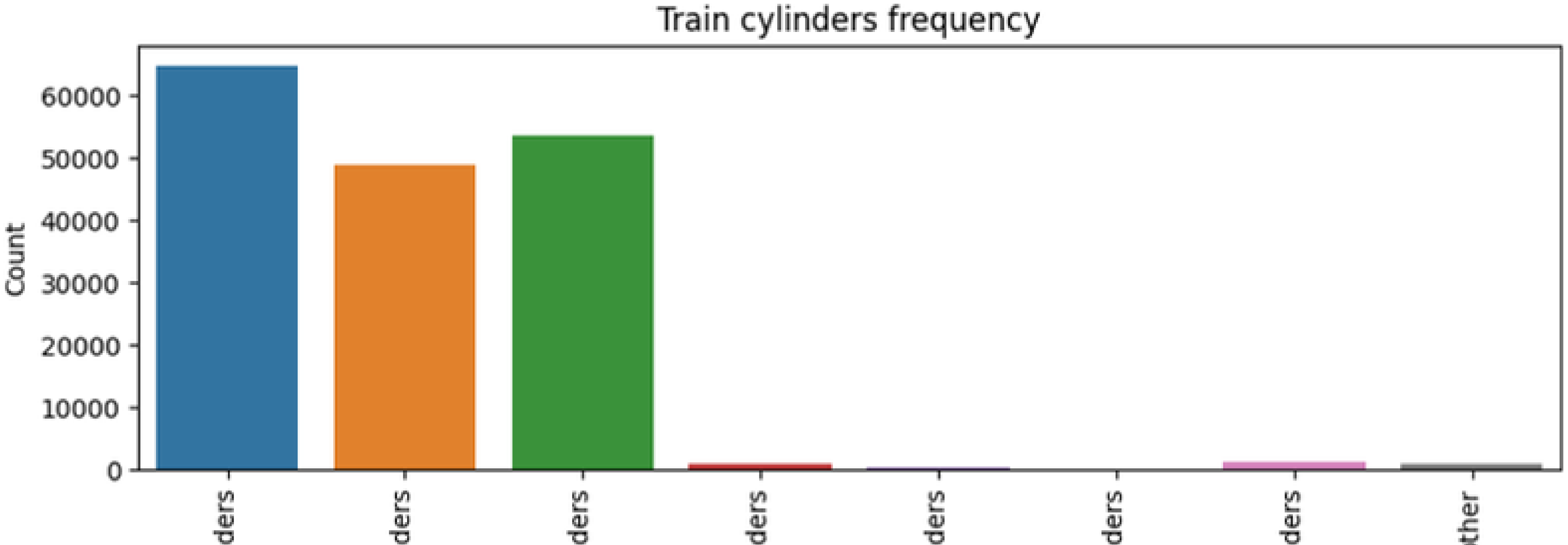
```
title_status
clean           275824
missing           5572
rebuilt           4924
salvage           2692
lien               981
parts only         136
Name: count, dtype: int64
```



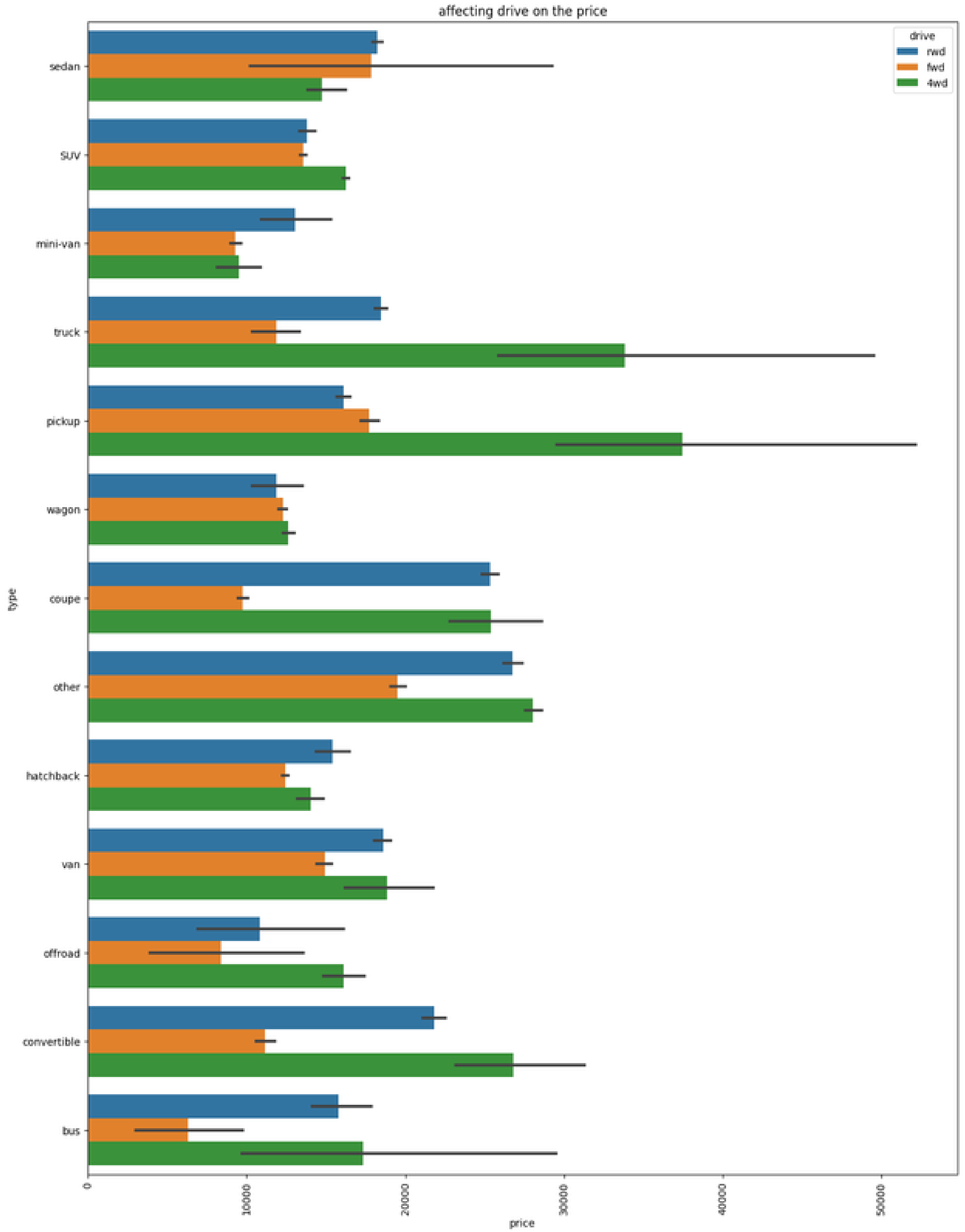
CYLINDERS

FILLING THE NULL VALUE BY ‘OTHER’ VALUE

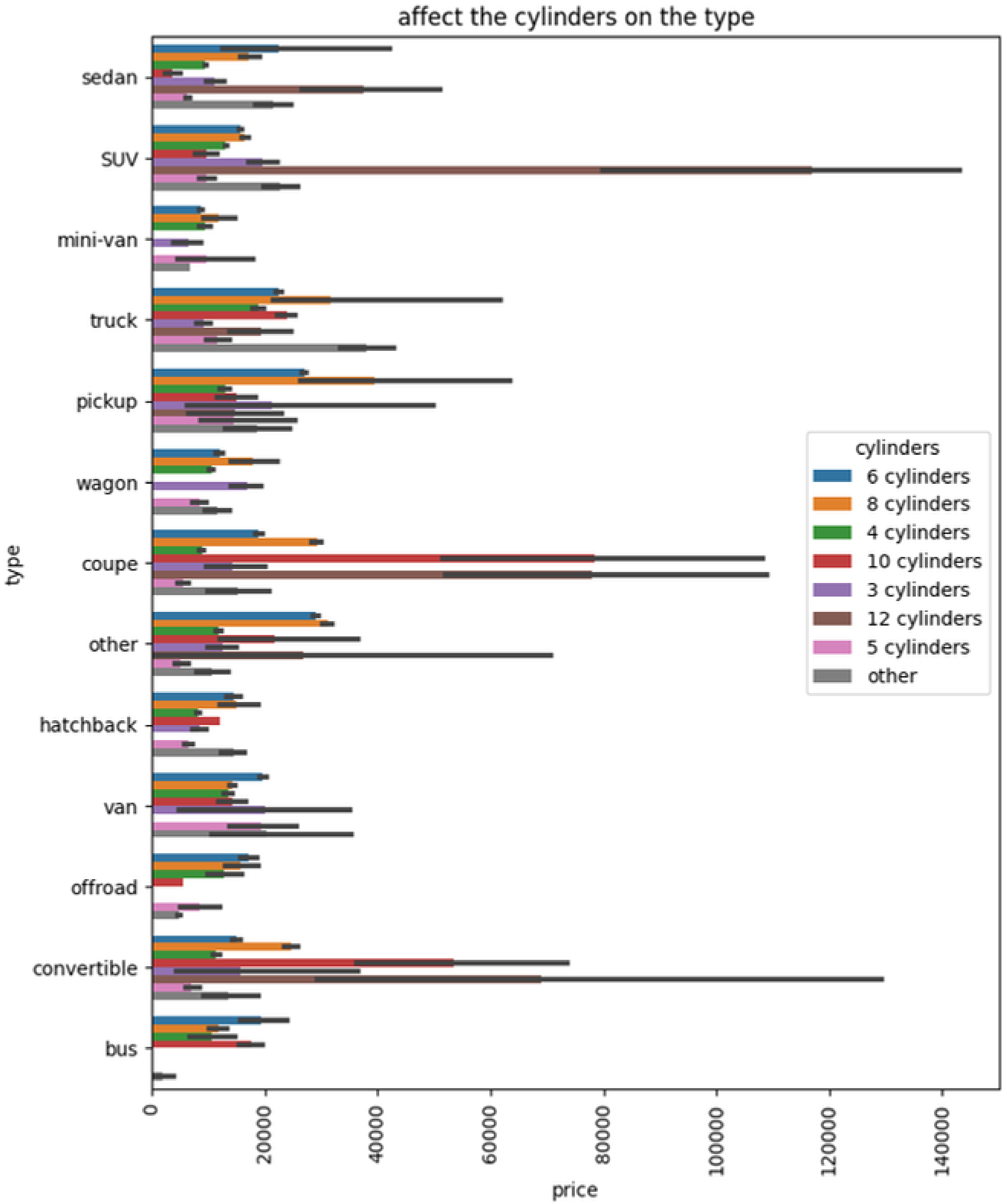




TYPE



affecting drive on the price
drive
rwd
fwd
4wd
sedan
SUV
mini-van
truck
pickup
wagon
coupe
other
hatchback
van
offroad
convertible
bus
type
0
10000
20000
30000
40000
50000
price

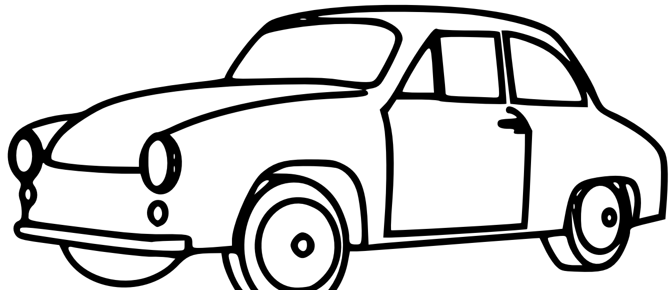


affect the cylinders on the type
cylinders
6 cylinders
8 cylinders
4 cylinders
10 cylinders
3 cylinders
12 cylinders
5 cylinders
other
sedan
SUV
mini-van
truck
pickup
wagon
coupe
other
hatchback
van
offroad
convertible
bus
type
0
20000
40000
60000
80000
100000
120000
140000
price

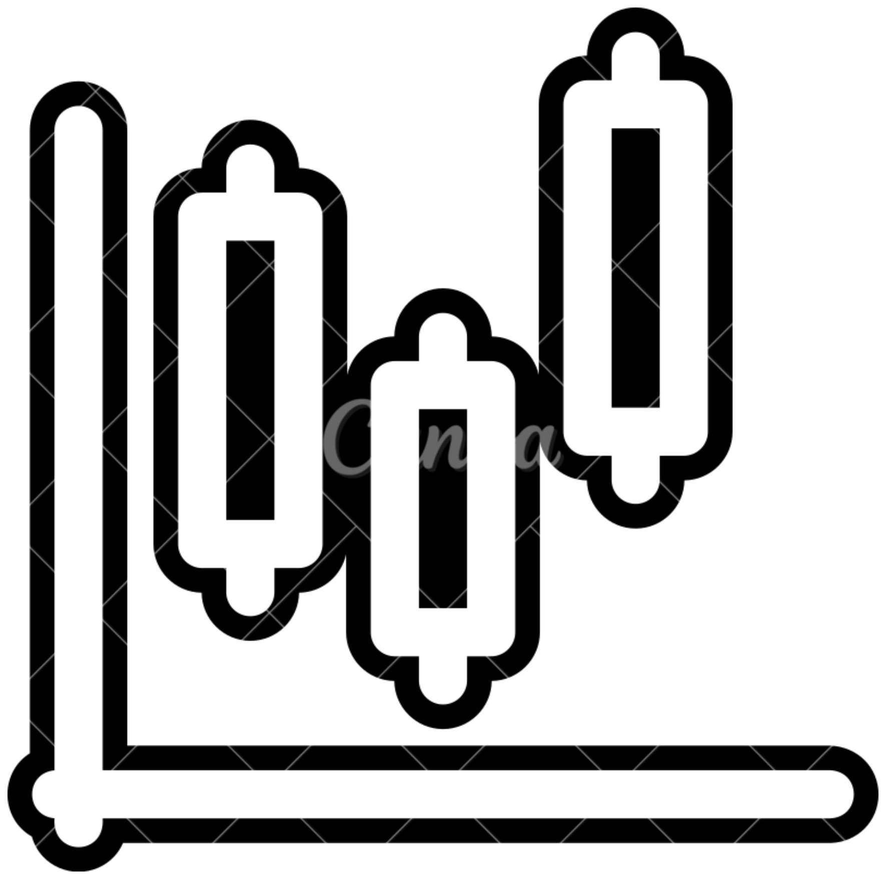
TYPE

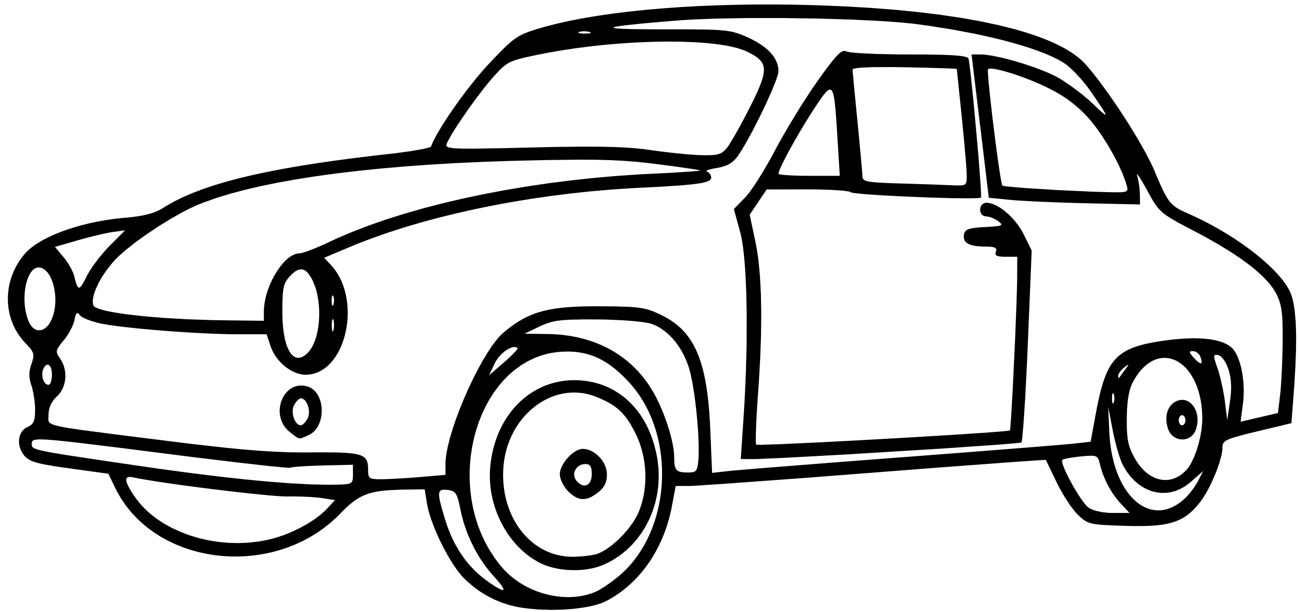
WE USE 'CYLINDERS', 'DRIVE‘ FEATURE TO FILL MISSING VALUE IN TYPE FEATURE.

By applying KNN algorithm we use search grid to find the best value for K and we obtain:
Best Parameters: {'n_neighbors': 500}

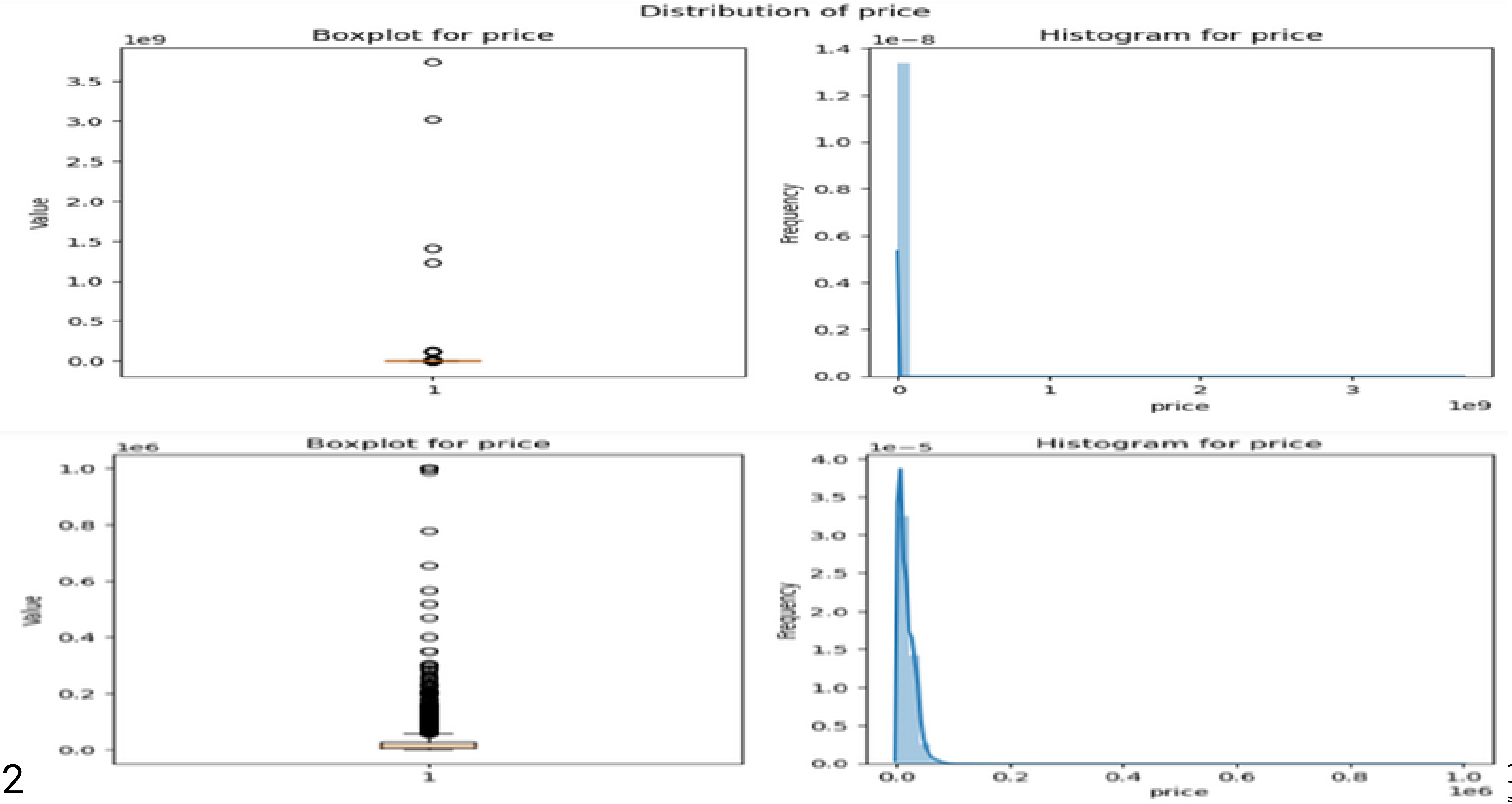


OUTLIERS

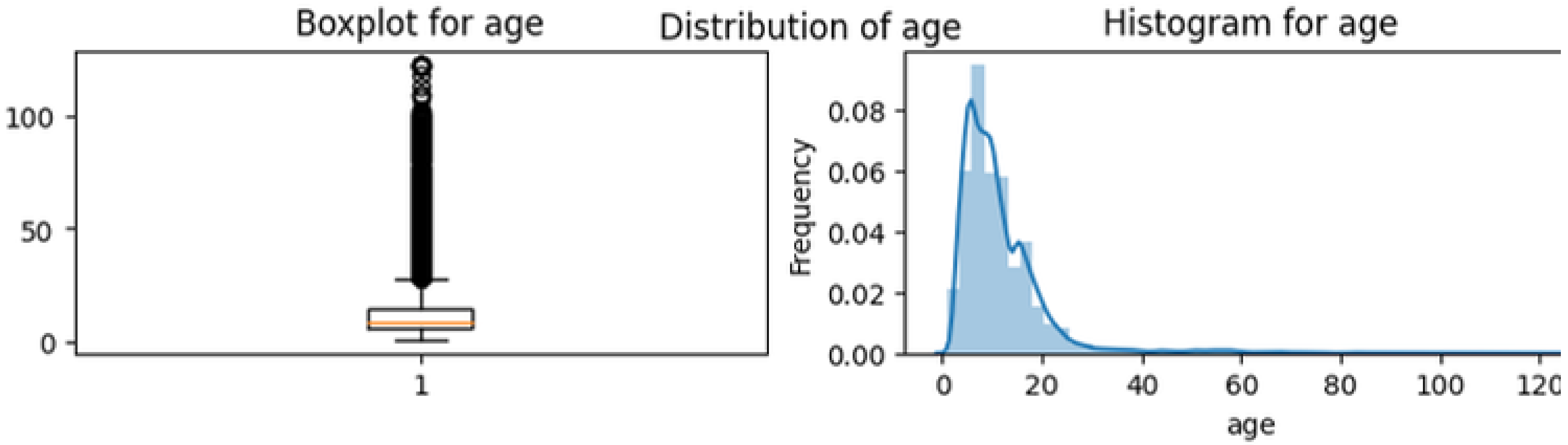




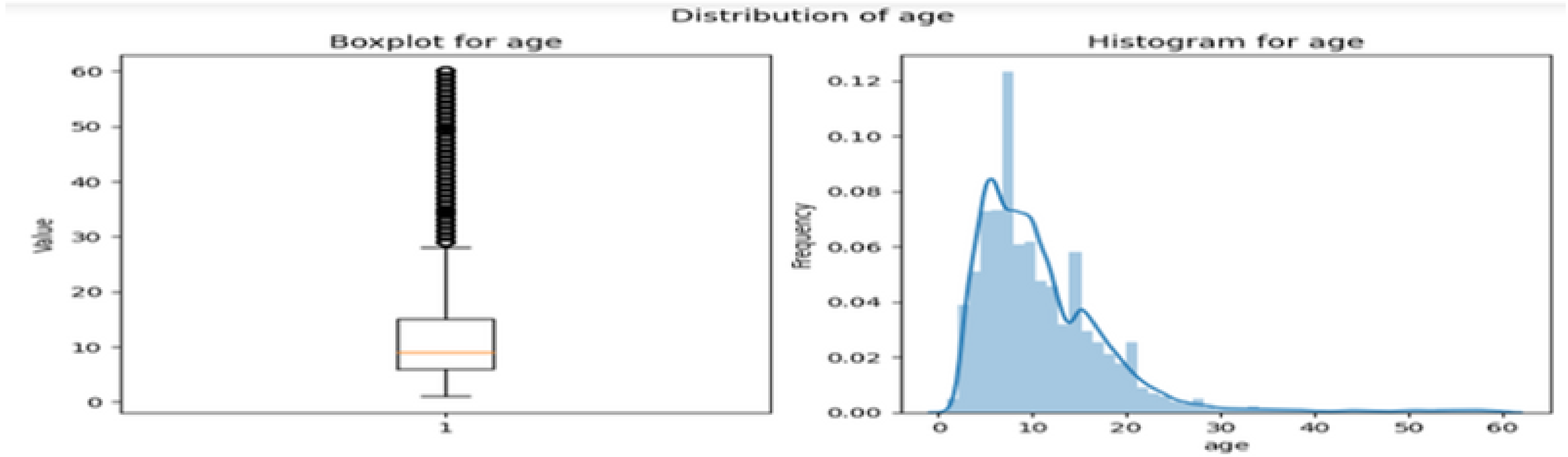
we have dropped only the data that have a price greater than 1,000,000



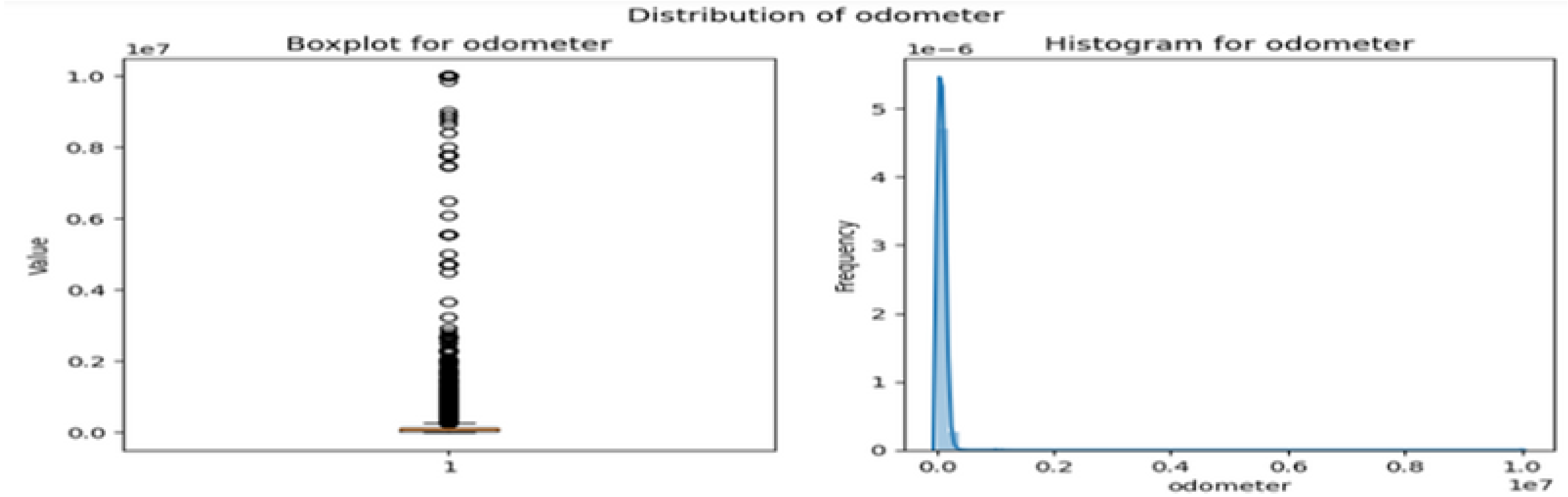
AGE

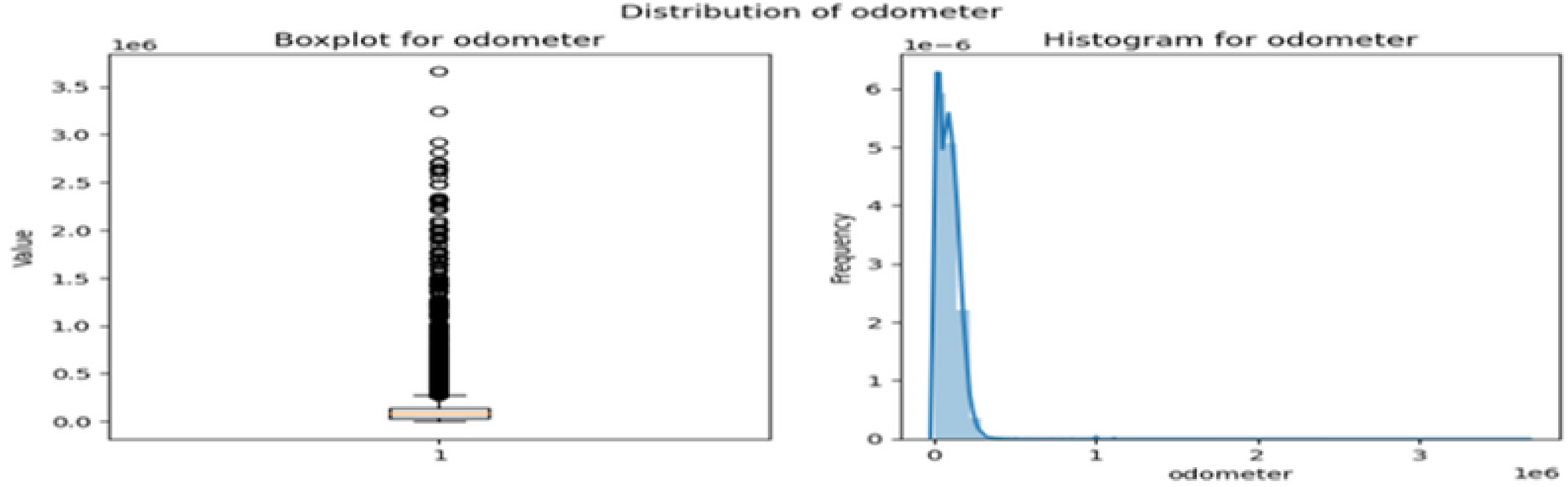


We handle the outlier in age by by put the value above 60 to mean and apply it on model



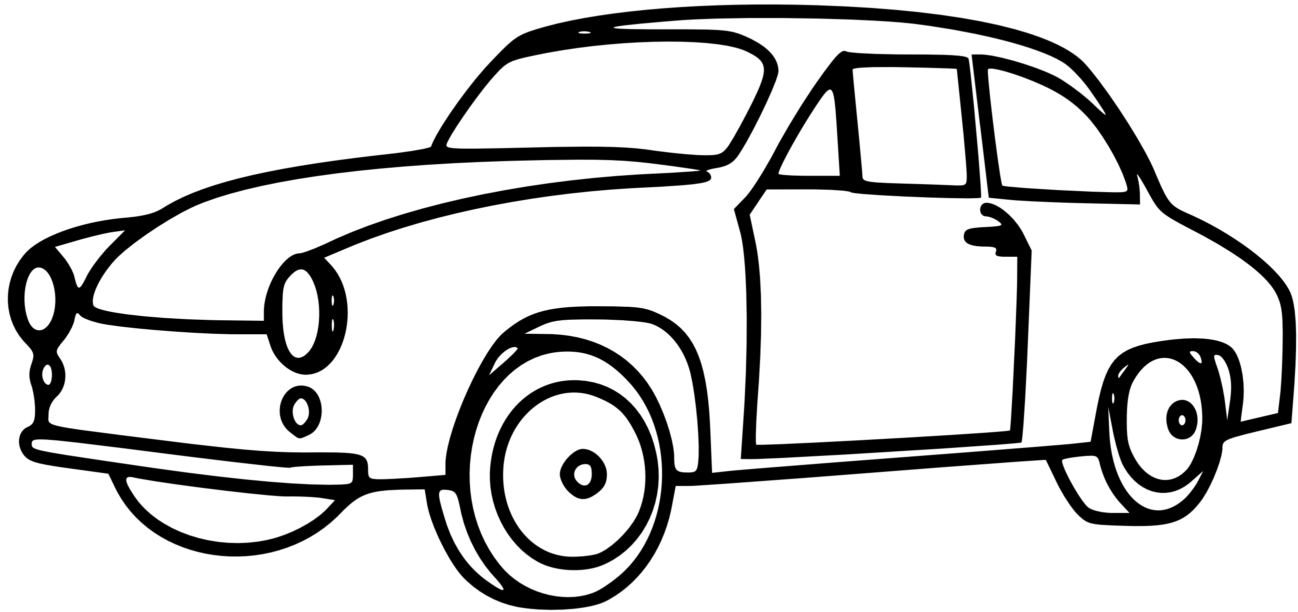
‘odometer’ feature has mean=97636.81368726927 and median=85615.0
we fill the value in odometer which has values > 0.4e7 with mean.



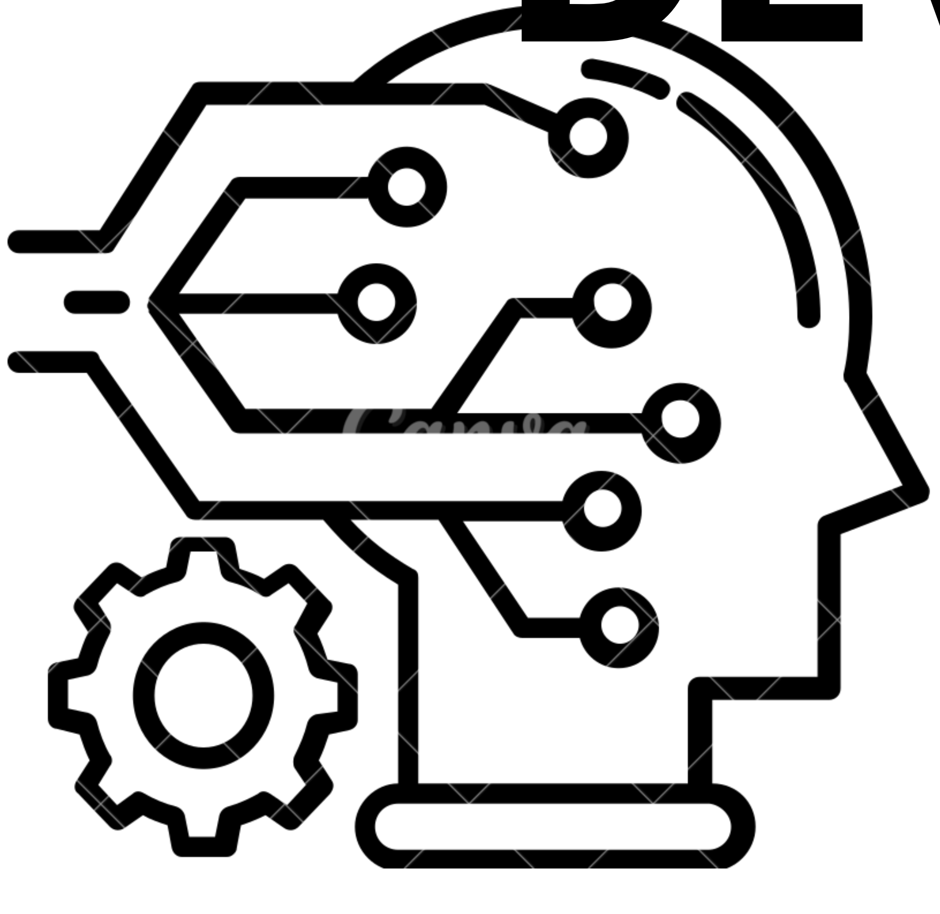


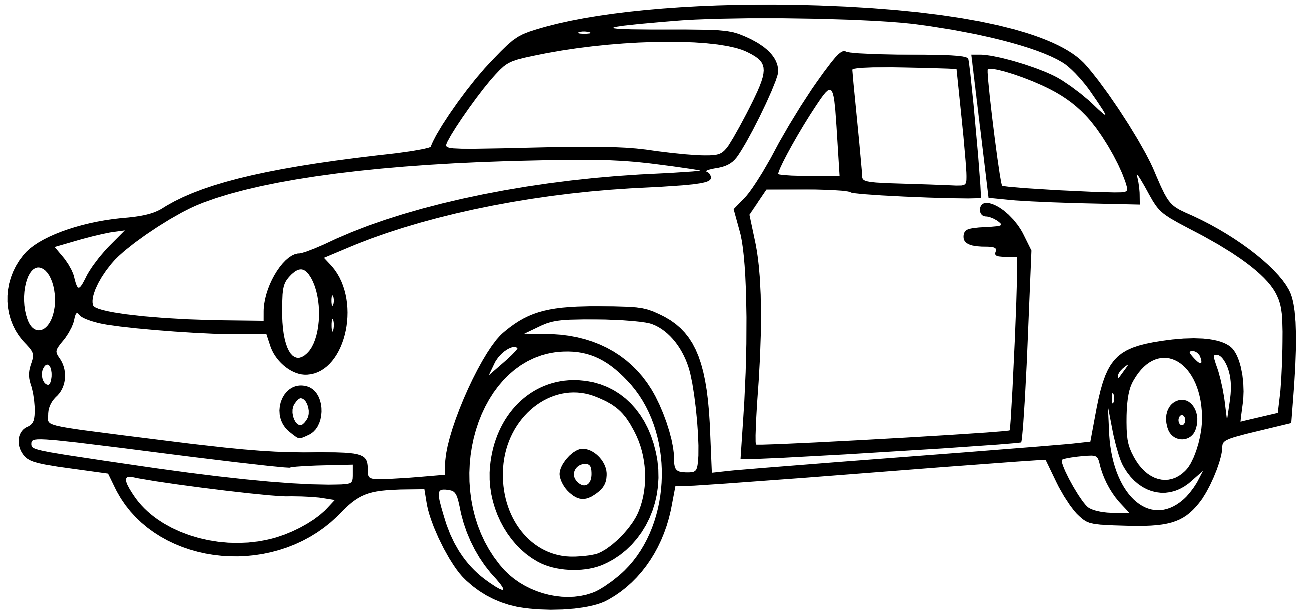
ENCODING DATA.
USING LDA

SCALING DATA USING
STANDARD SCALAR.



MODEL DEVELOPMENT





Following the data cleaning and preprocessing procedures, we tested various machine learning models to identify the most effective approach for our analysis

POLYNOMIAL REGRESSION (DEGREE =2 & 3)

BAGGING DECISION TREE

KNN (K = 1…….10, 20, 50, 100)

XGBOOST REGRESSOR

RF (WITH DIFFERENT HYPERPARAMETERS)

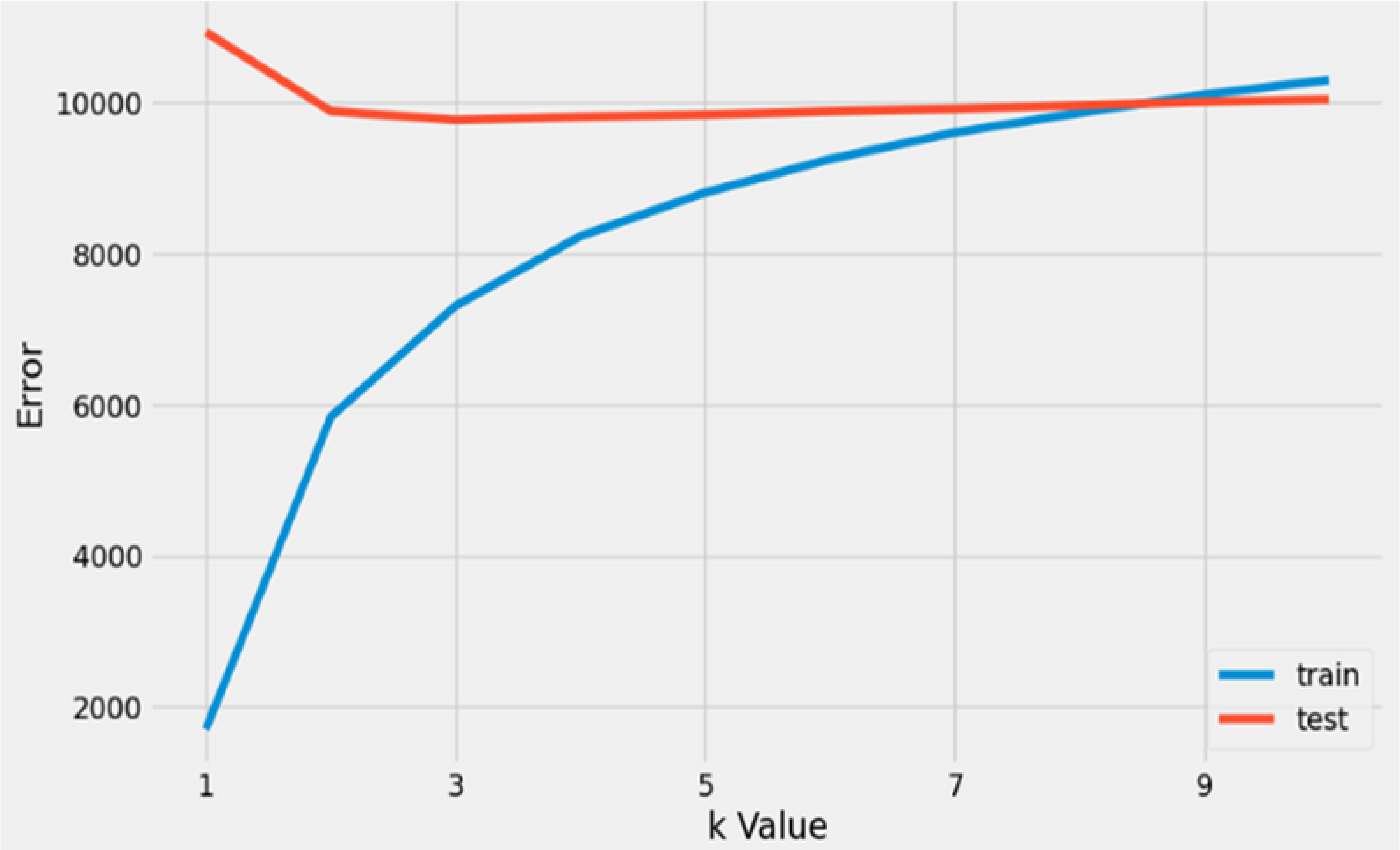
EXTRATREESREGRESSOR

POLYNOMIAL REGRESSION

Hyperparameter(degree)	R2_score for train	R2_score for validation	Kaggle score
2	.3775	.3086	We don't try it b/c the score is very low
3	0.4180	0.3960	We don't try it

KNN

WE TRAINED THE KNN WITH DIFFERENT K. THE LEARNING CURVE EXPLAIN THE



KNN

Hyperparamete r(K)	R2_score for train	R2_score for validation	Kaggle score
2	0.8739	0.5927	Don't run
3	0.8025	0.6021	Don't run

BAGGING REGRESSOR (USING DECISION TREE REGRESSOR)

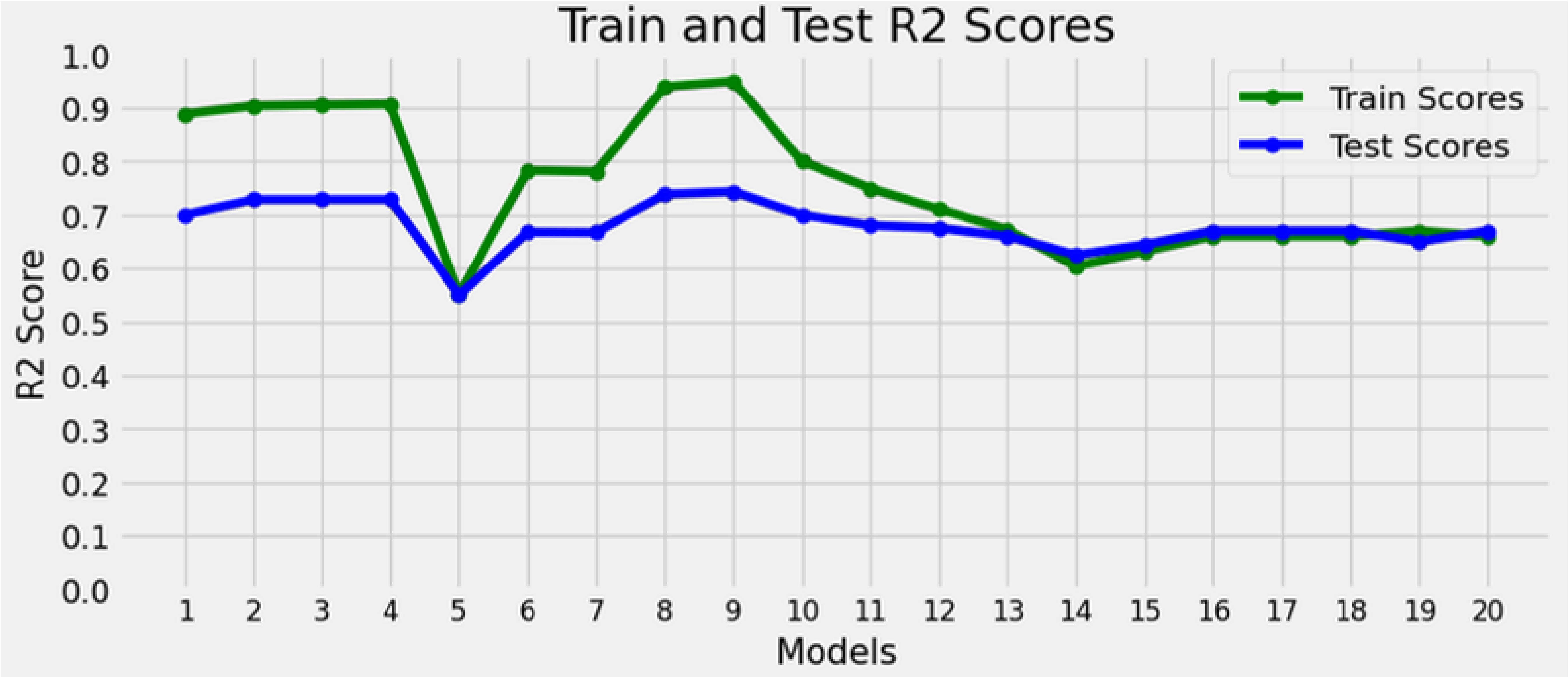
HYPERPARAMETER(DEGREE)	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
base_estimator=DecisionTreeRegressor(max_depth=20) n_estimators=50 random_state=1 max_samples=1.0 max_features=1.0	0.9026	0.7237	0.00038

RANDOM FOREST REGRESSOR

Hyper parameters	1	2	3	4	5	6	7	8	9	10
N_estimators	10	50	100	180	180	100	50	50	50	50
Max_depth	20	20	20	20	10	15	15	25	30	20
Min_samples_leaf	1	1	1	1	1	1	1	1	1	3
Random_state	0	0	0	0	0	0	0	0	0	0
R2_score train	0.889	0.904	0.906	0.9072	0.5503	0.7832	0.781	0.94	0.95	0.8
R2_score validation	0.7	0.729	0.729	0.729	0.55	0.66	0.667	0.739	0.744	0.7
KAGGLE SCORE		0.0004								0.0004

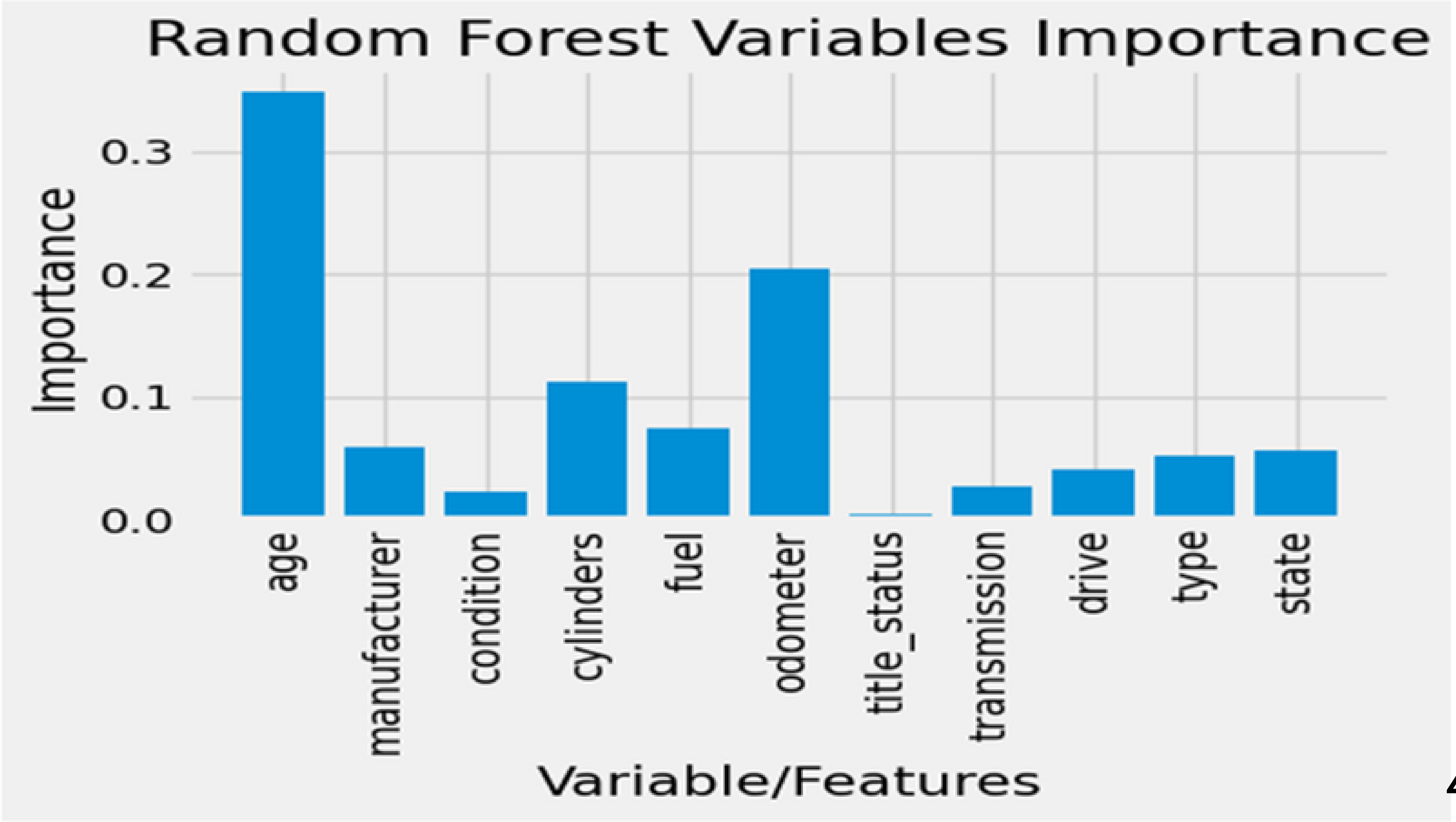
DOM FOREST REGRESSOR

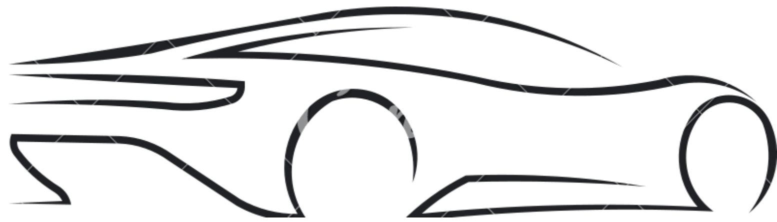
Hyper parameters	11	12	13	14	15	16	17	18	19	20
N_estimators	50	50	50	50	50	50	50	50	50	100
Max_depth	20	20	20	20	20	20	20	20	20	20
Min_leaf_samples	5	7	10	20	15	10	10	10	10	10
Random_state	00	0	0	0	0	1	10	40	100	0
R2_score train	0.75	0.711	0.677	0.604	0.632	0.66	0.66	0.66	0.67	0.66
0.75R2_score validation	0.68	0.675	0.6598	0.625	0.644	0.67	0.67	0.67	0.65	0.67
KAGGLE SCORE			0.00044							


Train and Test R2 Scores
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
R2 Score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Models
Train Scores
Test Scores

FEATURE IMPORTANCE FOR RFR

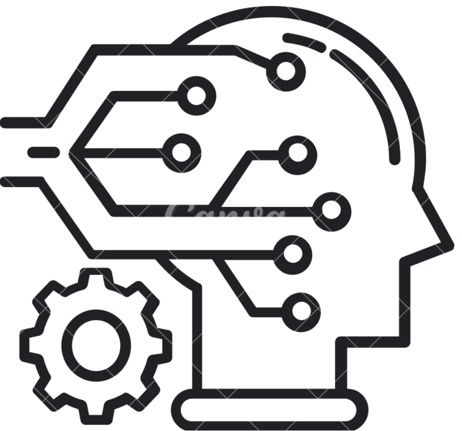
title_status has lowest importance what happened if we drop it?





DROP TITLE_STATUS

RFR	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
best one (trial number 13)	0.6713	0.6582	0.00044



XGBOOST REGRESSOR

HYPERPARAMETERS	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
n_estimators=500, max_depth=10, subsample=0.5	.9632	.6725	0.00008

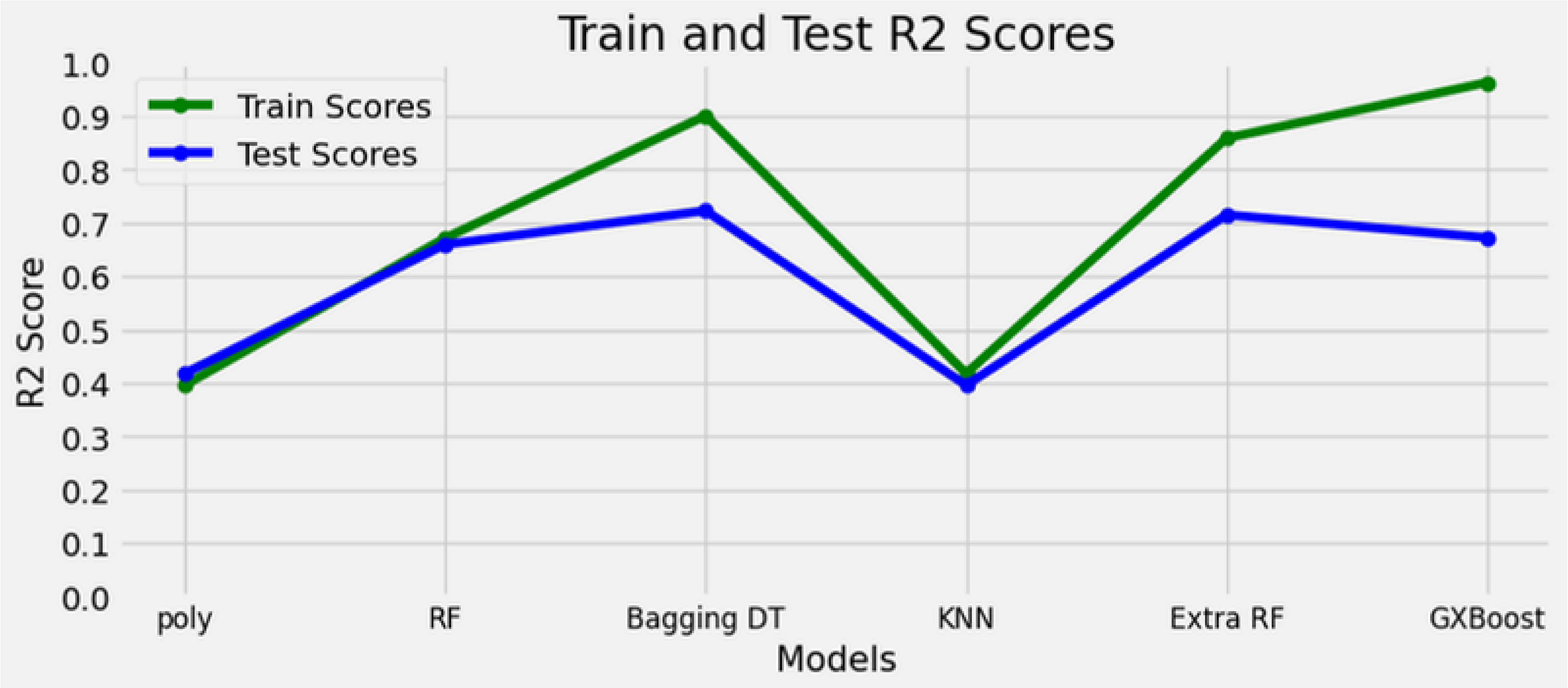
XGBOOST REGRESSOR

BEST HYPERPARAMETERS	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
'learning_rate': 0.5, 'max_depth': 7, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 1,	0. 7641	0. 6397	0.00018

EXTRATREESREGRESSOR

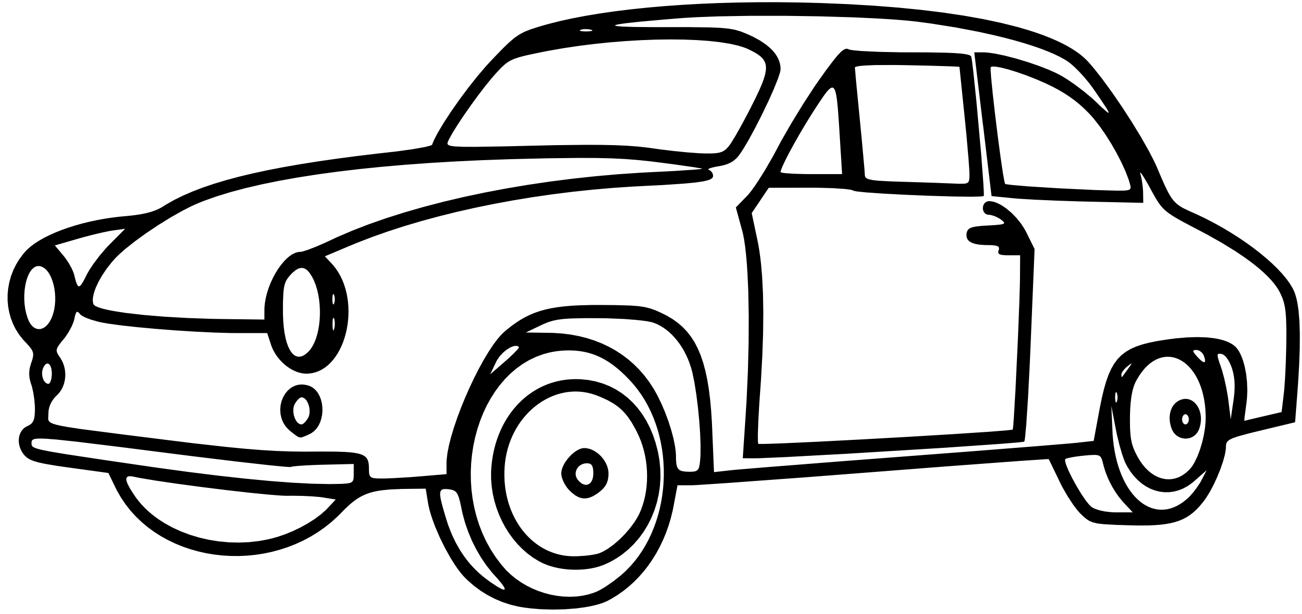
HYPERPARAMETERS	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
n_estimators=50, max_features=4000, max_depth=30, min_samples_leaf=3, n_jobs=-1	.859	.7153	0.00038
n_estimators=50, random_state=0, max_depth=20, min_samples_leaf=10	.6472	.6384	0.0003

SCORE FOR THE VARIOUS MODEL



OTHER TRIALS:

RFR	R2_SCORE FOR TRAIN	R2_SCORE FOR VALIDATION	KAGGLE SCORE
Handling the outliers by the median rather than the mean	0.6717	06598	0.00044
Dropping the outliers that has price greater than Q3 (26500.0) and lower than Q1 (5991.0)	0.796	0.7258	0.00012



THANK YOU