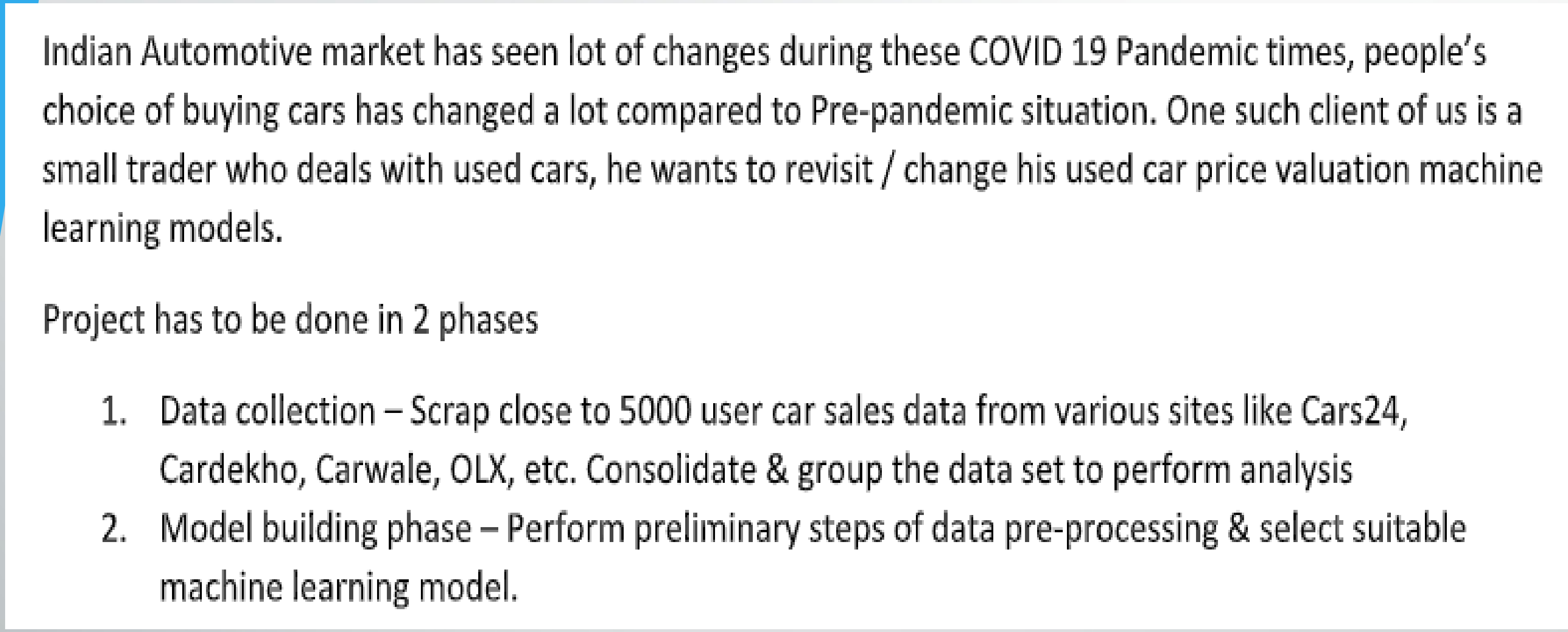
CAR PRICE PREDICTION

Amritesh Kumar

Dec 2021

ABSTRACT



Indian Automotive market has seen lot of changes during these COVID 19 Pandemic times, people's choice of buying cars has changed a lot compared to Pre-pandemic situation. One such client of us is a small trader who deals with used cars, he wants to revisit / change his used car price valuation machine learning models.

Project has to be done in 2 phases

1. Data collection – Scrap close to 5000 user car sales data from various sites like Cars24, Cardekho, Carwale, OLX, etc. Consolidate & group the data set to perform analysis
2. Model building phase – Perform preliminary steps of data pre-processing & select suitable machine learning model.

EXPERIMENTAL SET UP

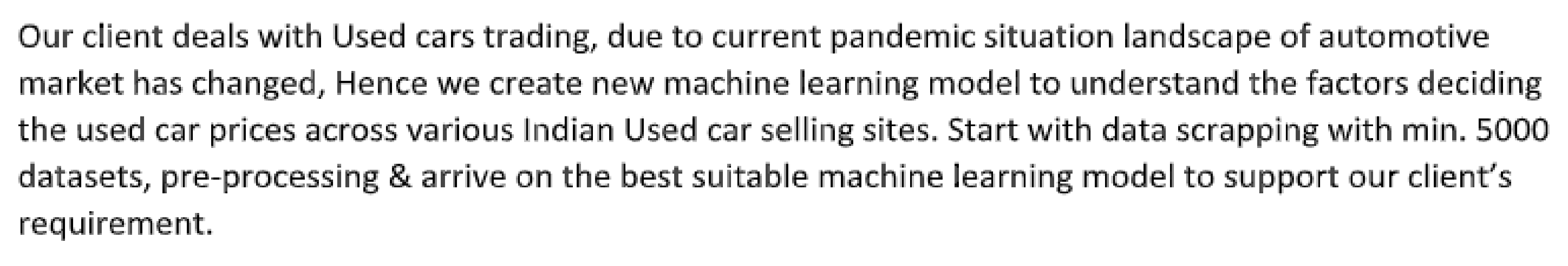
Hardware requirements:-

1. Processor — core i5 and above
2. RAM — 8 GB or above
3. SSD— 250 GB or above

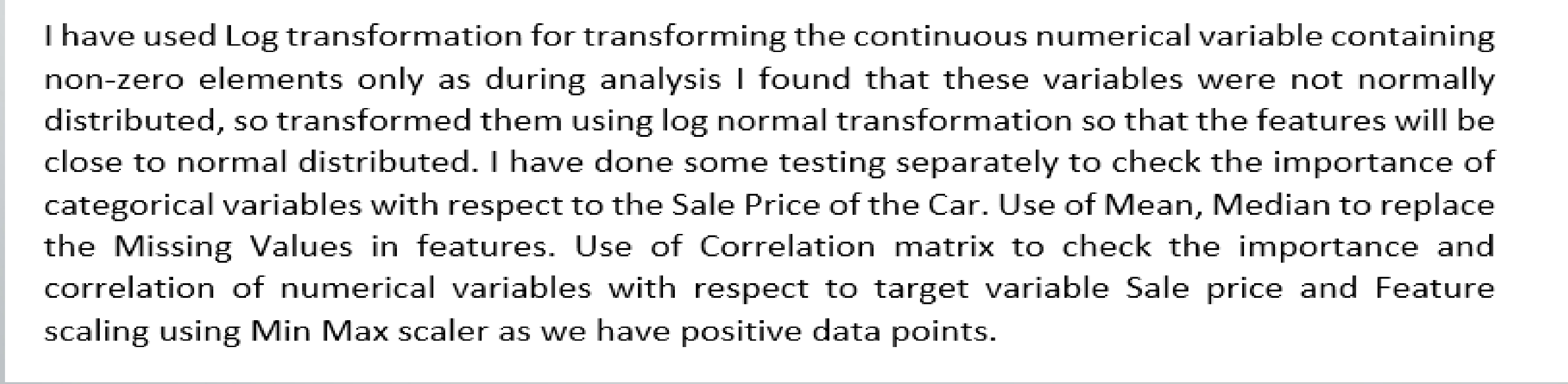
Software requirements:-

1. ANACONDA

INTRODUCTION



Our client deals with Used cars trading, due to current pandemic situation landscape of automotive market has changed, Hence we create new machine learning model to understand the factors deciding the used car prices across various Indian Used car selling sites. Start with data scrapping with min. 5000 datasets, pre-processing & arrive on the best suitable machine learning model to support our client’s requirement.



I have used Log transformation for transforming the continuous numerical variable containing non-zero elements only as during analysis I found that these variables were not normally distributed, so transformed them using log normal transformation so that the features will be close to normal distributed. I have done some testing separately to check the importance of categorical variables with respect to the Sale Price of the Car. Use of Mean, Median to replace the Missing Values in features. Use of Correlation matrix to check the importance and correlation of numerical variables with respect to target variable Sale price and Feature scaling using Min Max scaler as we have positive data points.

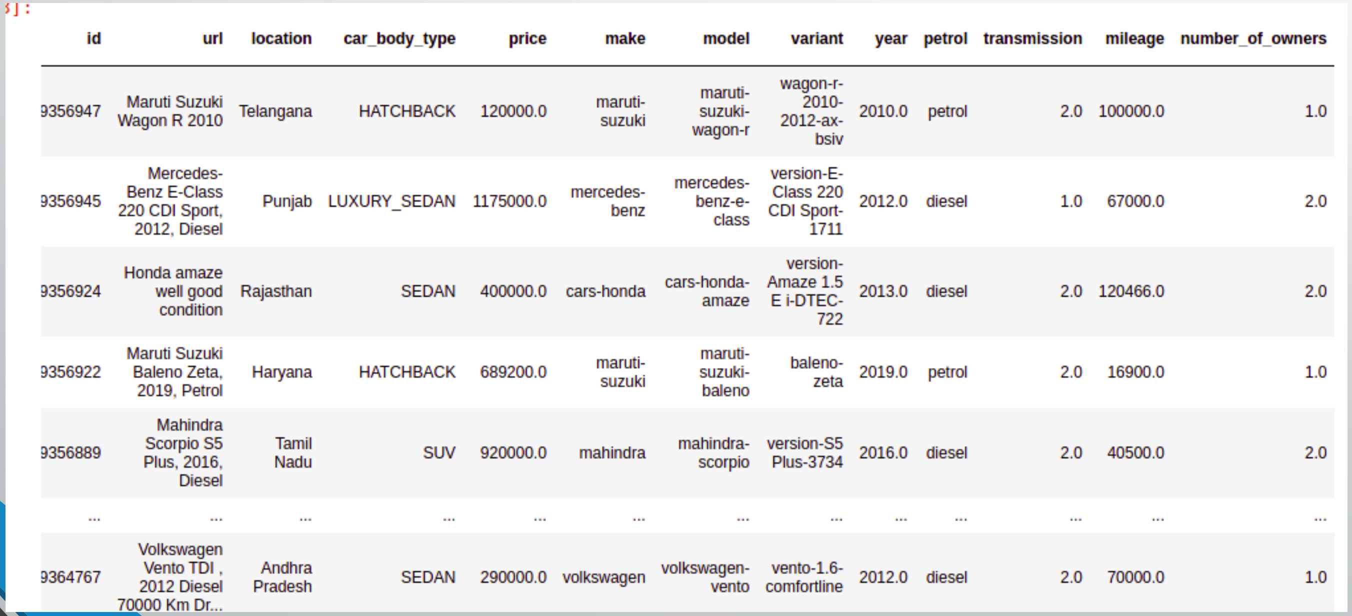
DATA PREPRATION

With the help of Pandas Library, We will upload our data to Jupyter Notebook.

Once our data is uploaded with the help of predefined method (i.e. read_csv) we can read data for further processing.

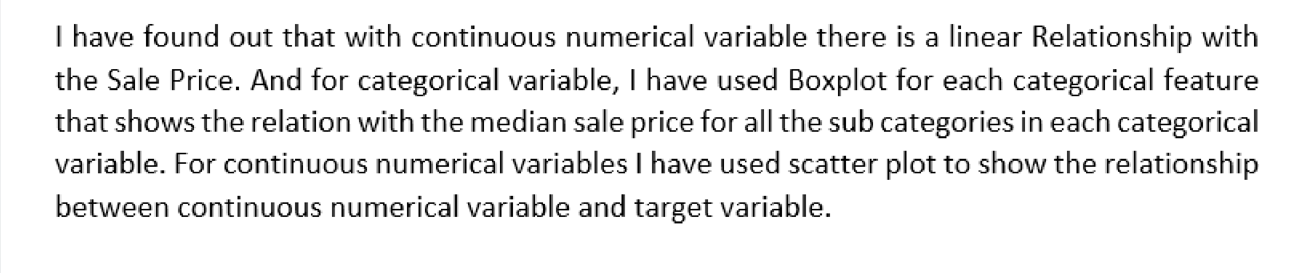
We have two type of variables in the data:-

- Dependent Variable
- Independent Variable



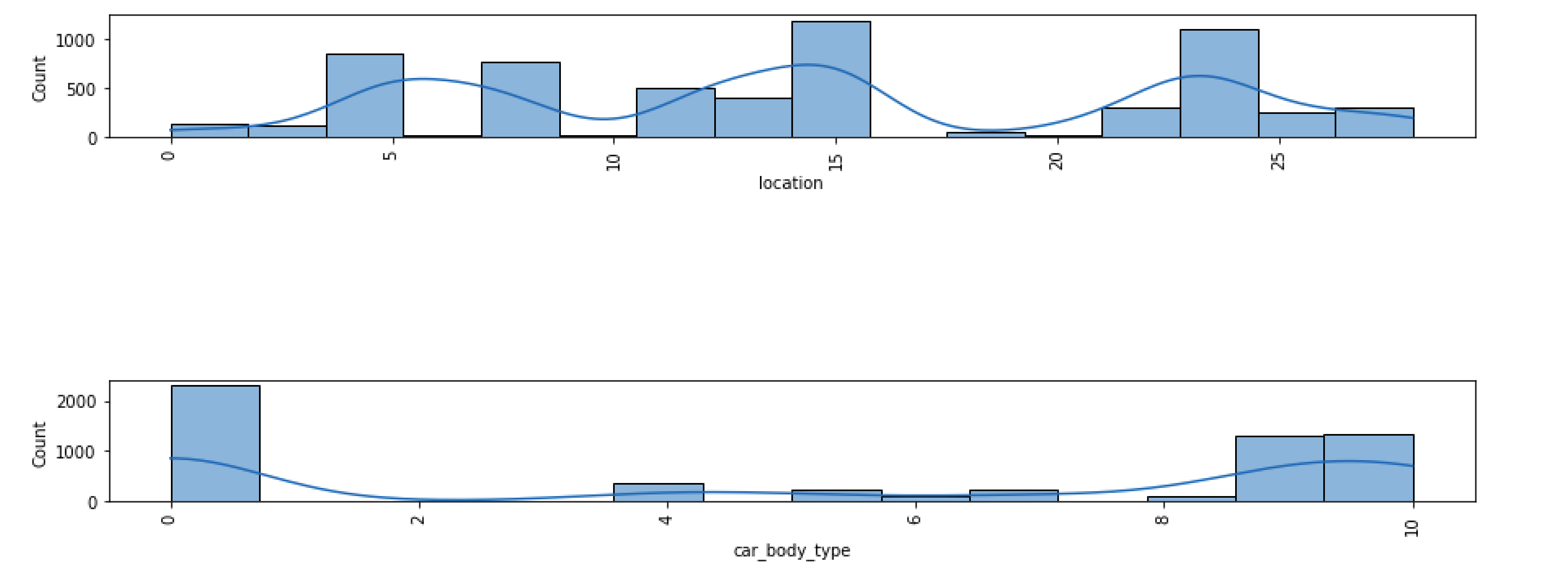
id	url	location	car_body_type	price	make	model	variant	year	petrol	transmission	mileage	number_of_owners
9356947	Maruti Suzuki Wagon R 2010	Telangana	HATCHBACK	120000.0	maruti-suzuki	maruti-suzuki-wagon-r	wagon-r-2010-2012-ax-bsiv	2010.0	petrol	2.0	100000.0	1.0
9356945	Mercedes-Benz E-Class 220 CDI Sport, 2012, Diesel	Punjab	LUXURY_SEDAN	1175000.0	mercedes-benz	mercedes-benz-e-class	version-E-Class 220 CDI Sport-1711	2012.0	diesel	1.0	67000.0	2.0
9356924	Honda amaze well good condition	Rajasthan	SEDAN	400000.0	cars-honda	cars-honda-amaze	version-Amaze 1.5 E i-DTEC-722	2013.0	diesel	2.0	120466.0	2.0
9356922	Maruti Suzuki Baleno Zeta, 2019, Petrol	Haryana	HATCHBACK	689200.0	maruti-suzuki	maruti-suzuki-baleno	baleno-zeta	2019.0	petrol	2.0	16900.0	1.0
9356889	Mahindra Scorpio S5 Plus, 2016, Diesel	Tamil Nadu	SUV	920000.0	mahindra	mahindra-scorpio	version-S5 Plus-3734	2016.0	diesel	2.0	40500.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...
9364767	Volkswagen Vento TDI , 2012 Diesel 70000 Km Dr...	Andhra Pradesh	SEDAN	290000.0	volkswagen	volkswagen-vento	vento-1.6-comfortline	2012.0	diesel	2.0	70000.0	1.0

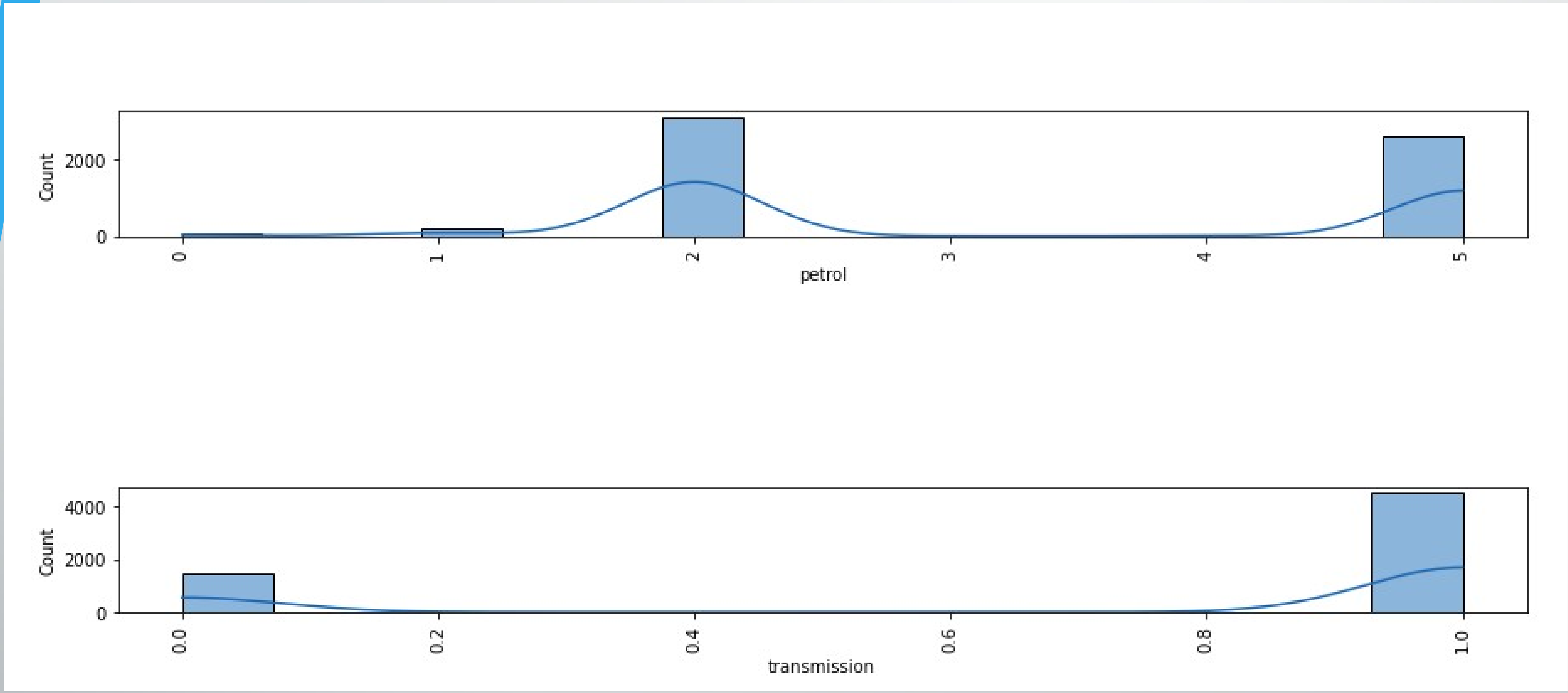
Car price is a dependent variable whereas all of the other elements are independent variables.

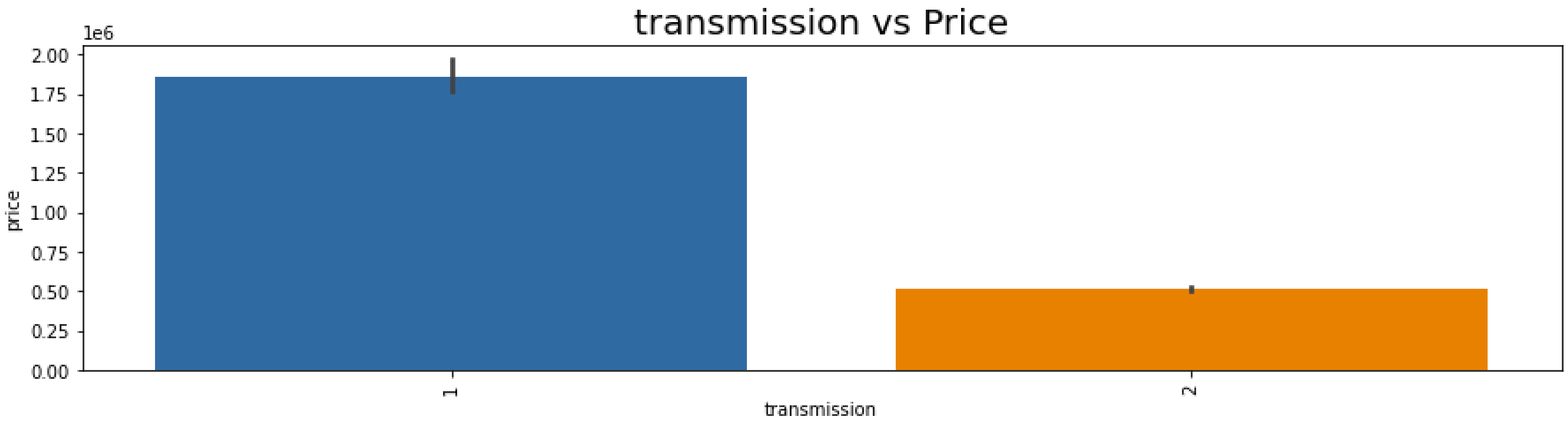


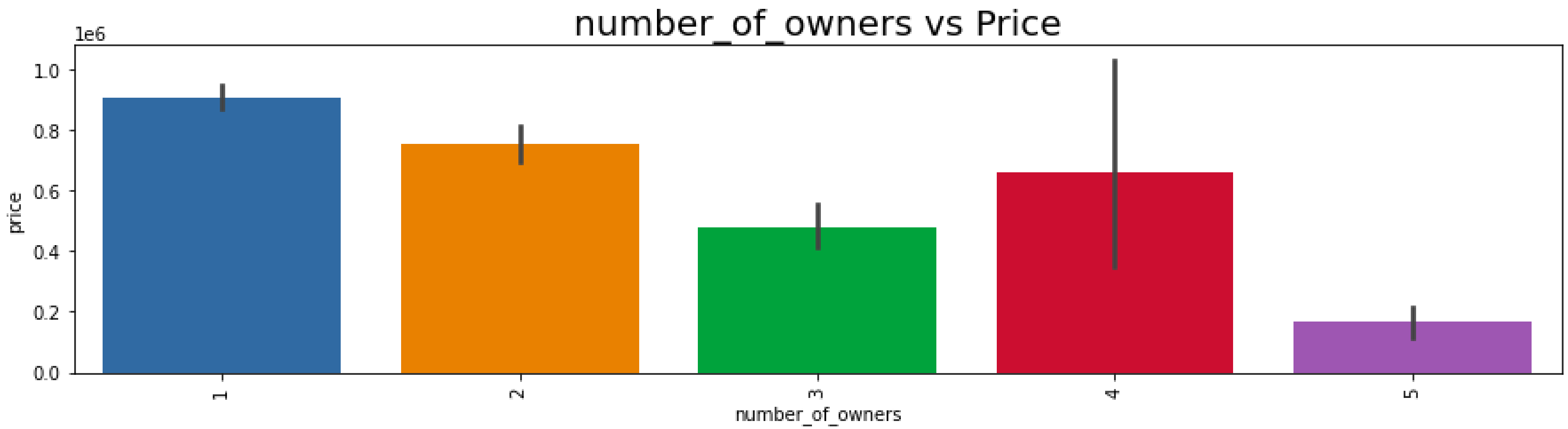
I have found out that with continuous numerical variable there is a linear Relationship with the Sale Price. And for categorical variable, I have used Boxplot for each categorical feature that shows the relation with the median sale price for all the sub categories in each categorical variable. For continuous numerical variables I have used scatter plot to show the relationship between continuous numerical variable and target variable.

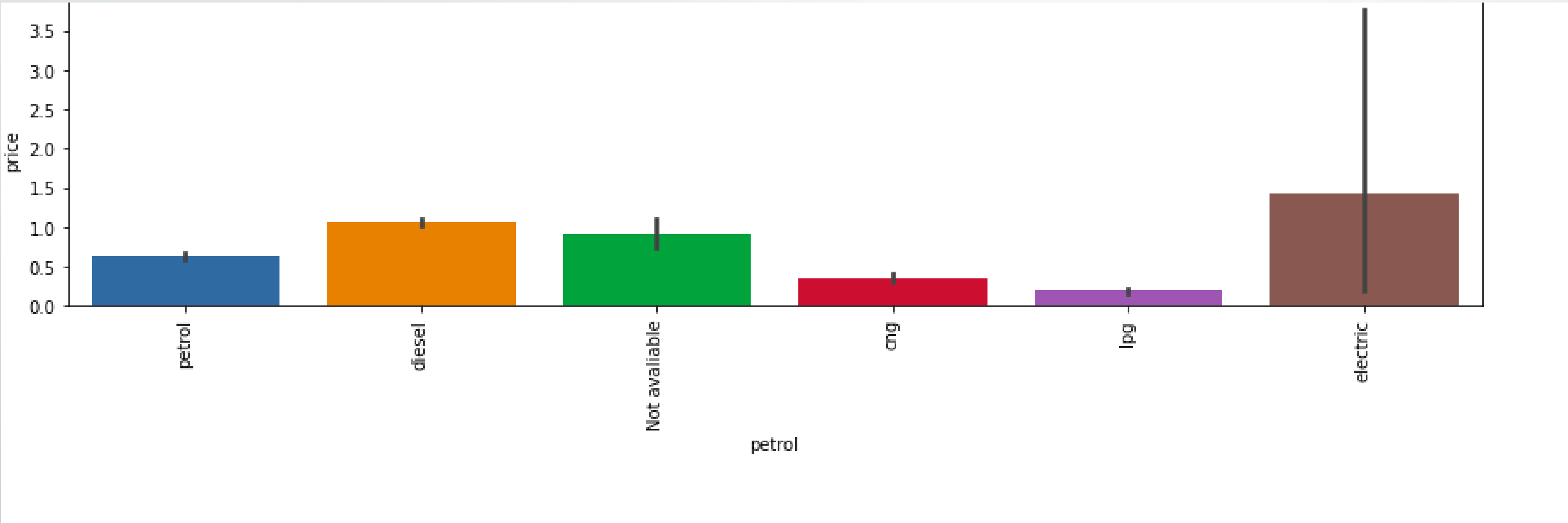
VISUALIZATION

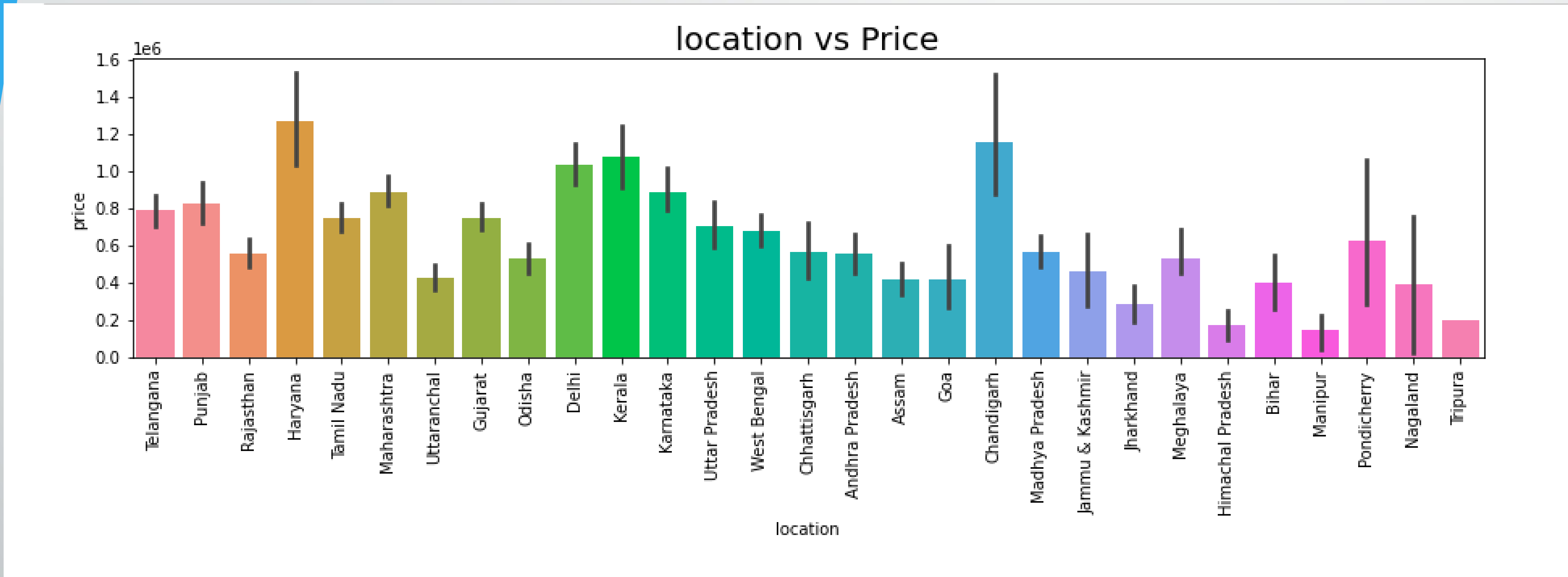


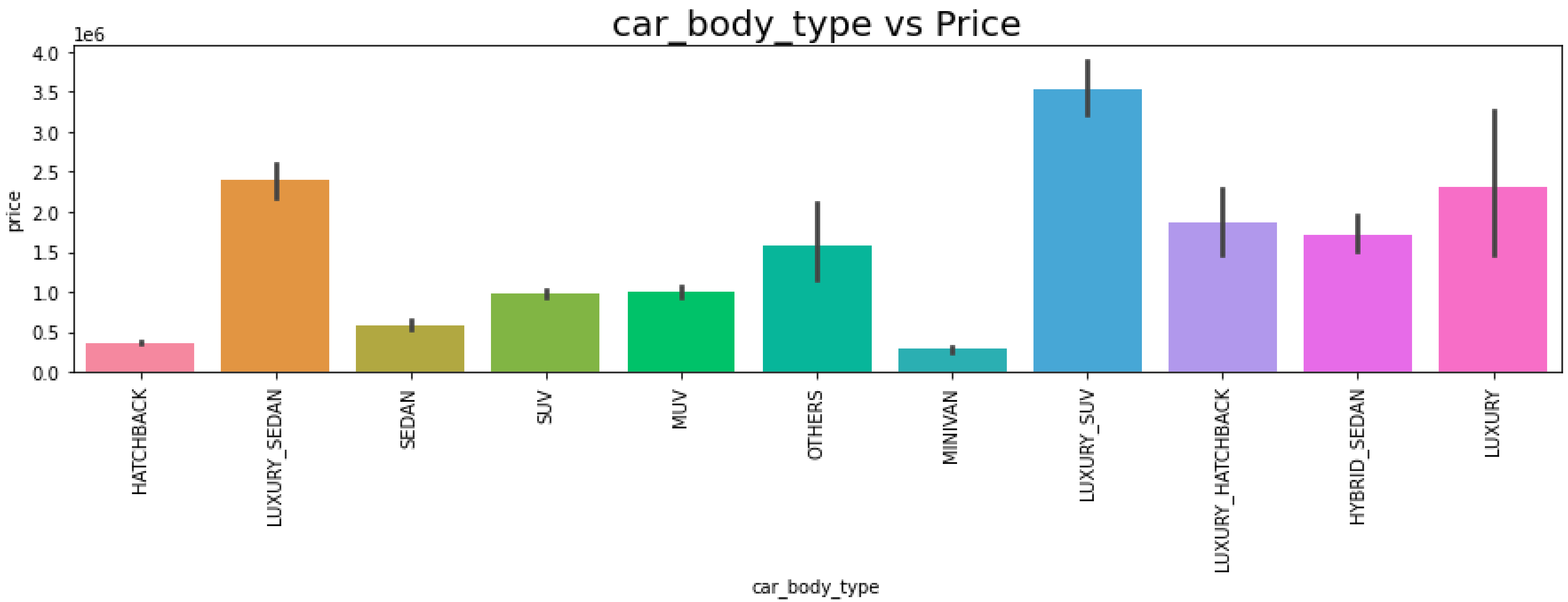

Count
2000
0
0
1
2
3
4
5
petrol
Count
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
transmission

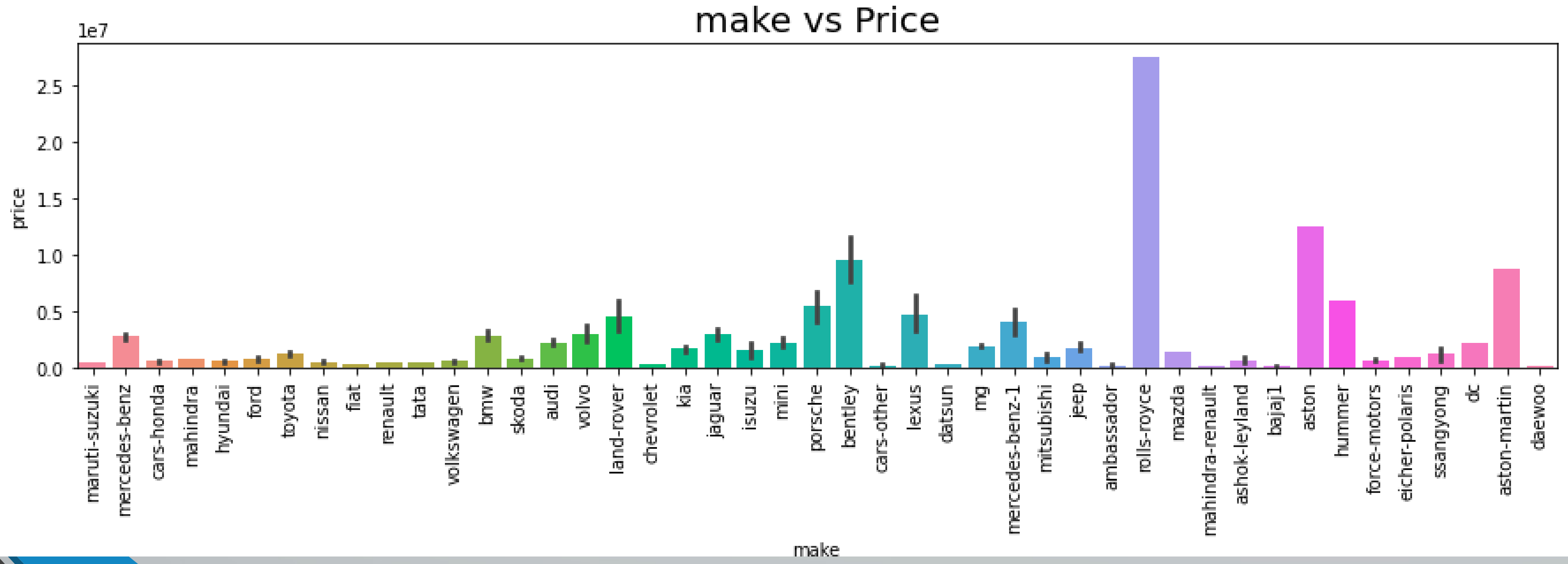

transmission vs Price
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
price
1
2
transmission


number_of_owners vs Price
1e6
1.0
0.8
0.6
0.4
0.2
0.0
price
1
2
3
4
5
number_of_owners


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
price
petrol
diesel
Not avaliable
cng
lpg
electric
petrol


location vs Price
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
price
Telangana
Punjab
Rajasthan
Haryana
Tamil Nadu
Maharashtra
Uttaranchal
Gujarat
Odisha
Delhi
Kerala
Karnataka
Uttar Pradesh
West Bengal
Chhattisgarh
Andhra Pradesh
Assam
Goa
Chandigarh
Madhya Pradesh
Jammu & Kashmir
Jharkhand
Meghalaya
Himachal Pradesh
Bihar
Manipur
Pondicherry
Nagaland
Tripura
location


car_body_type vs Price
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
price
HATCHBACK
LUXURY_SEDAN
SEDAN
SUV
MUV
OTHERS
MINIVAN
LUXURY_SUV
LUXURY_HATCHBACK
HYBRID_SEDAN
LUXURY
car_body_type


make vs Price
1e7
2.5
2.0
1.5
1.0
0.5
0.0
price
maruti-suzuki
mercedes-benz
cars-honda
mahindra
hyundai
ford
toyota
nissan
fiat
renault
tata
volkswagen
bmw
skoda
audi
volvo
land-rover
chevrolet
kia
jaguar
isuzu
mini
porsche
bentley
cars-other
lexus
datsun
mg
mercedes-benz-1
mitsubishi
jeep
ambassador
rolls-royce
mazda
mahindra-renault
ashok-leyland
bajaj1
aston
hummer
force-motors
eicher-polaris
ssangyong
dc
aston-martin
daewoo
make

DATA PREPROCESSING

- The Complete data is divided in the ration of 70:30 for train and test respectively.
- There is lots of null values present in the data-set and there are some outliers present in the data=set which has been replace with padding and not available .
- Once our data is ready categorical variables are converted into the numeric form, which we can apply further on algorithms.
- I have dropped the some column since there is no correlation between output variable and with those columns.

EVALUTION PROCESS

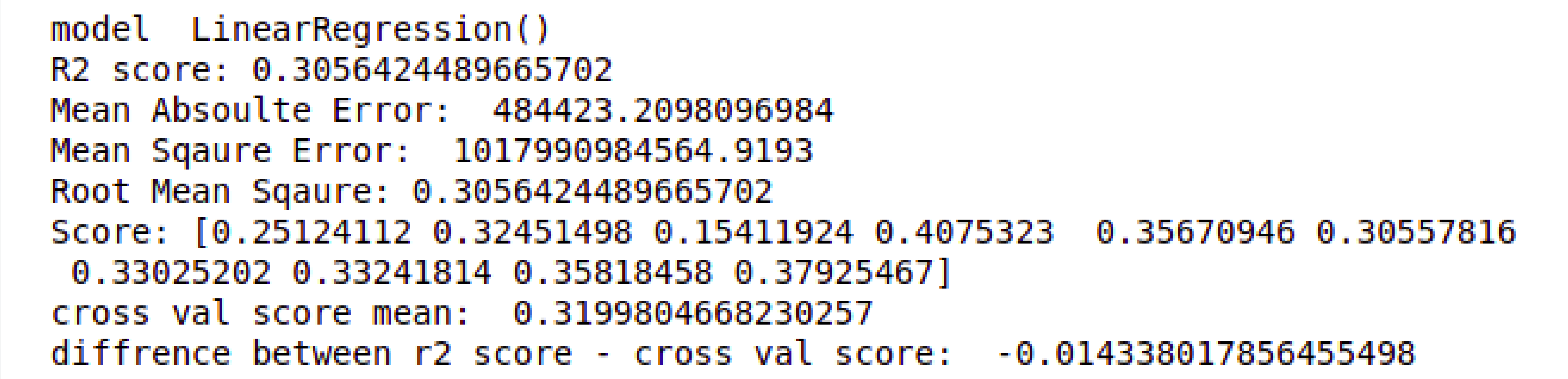
R2 Score:

- Adjusted R2 Score deals with additional independent variables.
- This R squared value of the r-square if our choice of independent variable wasn’t good (i.e. independent variable had no effect on dependent variable)
- Also the bias of R Square to not decrease is handled pretty well in this adjusted R Squared method.

Cross Validations:

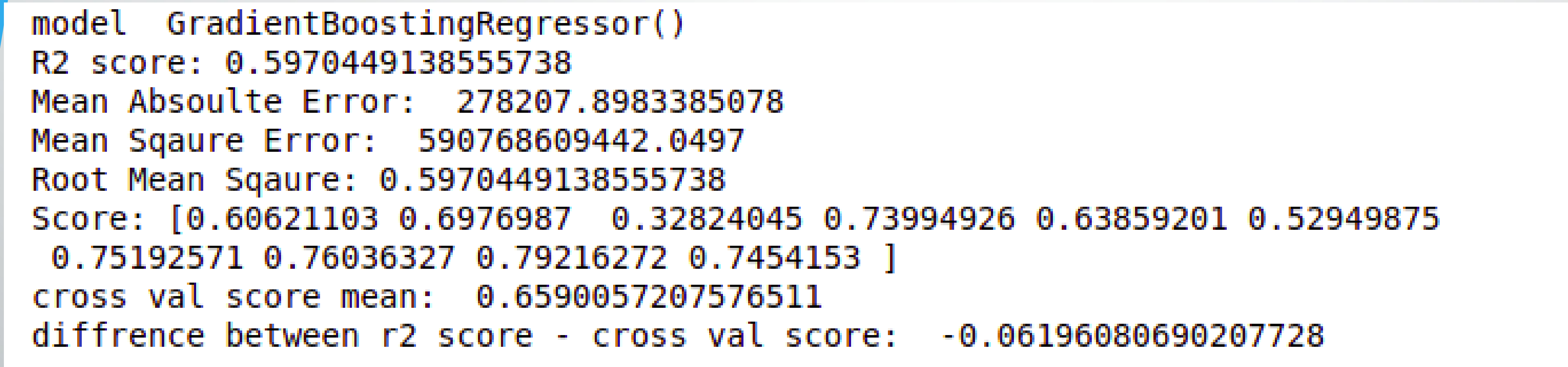
- K Fold cross validations , K = 5

Linear Regression



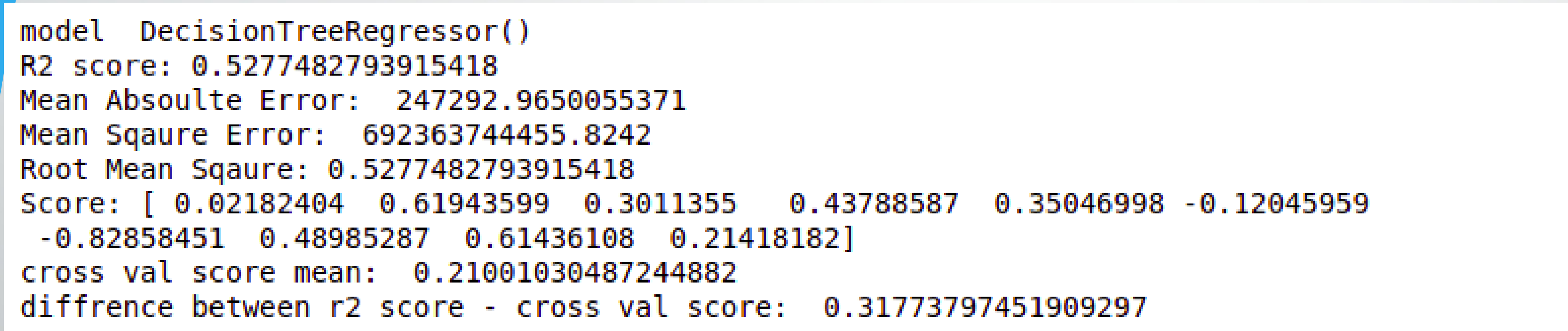
```
model  LinearRegression()
R2 score: 0.3056424489665702
Mean Absoulte Error:  484423.2098096984
Mean Sqaure Error:  1017990984564.9193
Root Mean Sqaure: 0.3056424489665702
Score: [0.25124112 0.32451498 0.15411924 0.4075323  0.35670946 0.30557816
 0.33025202 0.33241814 0.35818458 0.37925467]
cross val score mean:  0.3199804668230257
diffrence between r2 score - cross val score:  -0.014338017856455498
```

Gradient Boosting Regressor



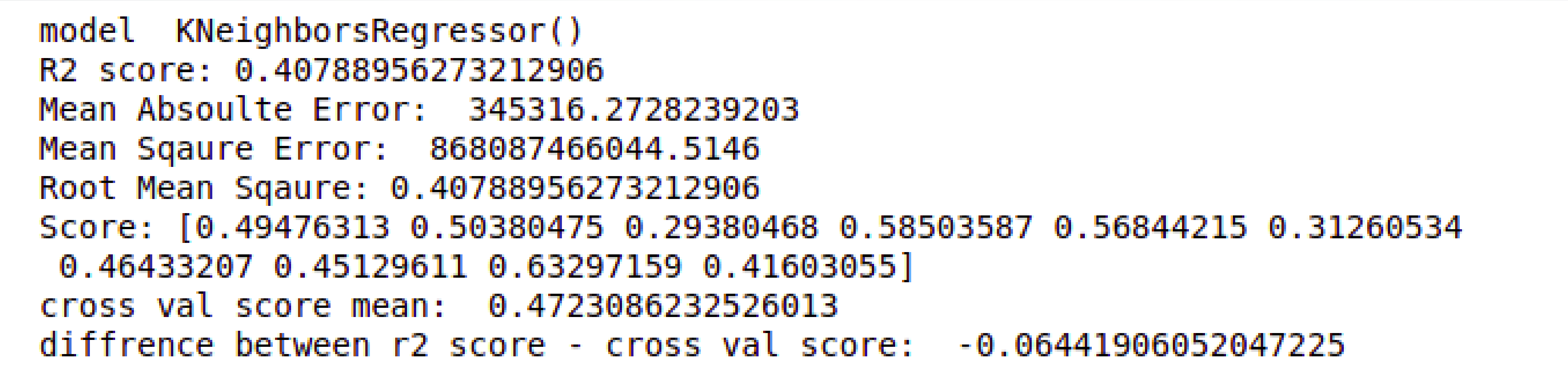
```
model  GradientBoostingRegressor()
R2 score: 0.5970449138555738
Mean Absoulte Error:  278207.8983385078
Mean Sqaure Error:  590768609442.0497
Root Mean Sqaure: 0.5970449138555738
Score: [0.60621103 0.6976987  0.32824045 0.73994926 0.63859201 0.52949875
 0.75192571 0.76036327 0.79216272 0.7454153 ]
cross val score mean:  0.6590057207576511
diffrence between r2 score - cross val score:  -0.0619608069020772 8
```

Decision Tree Regressor



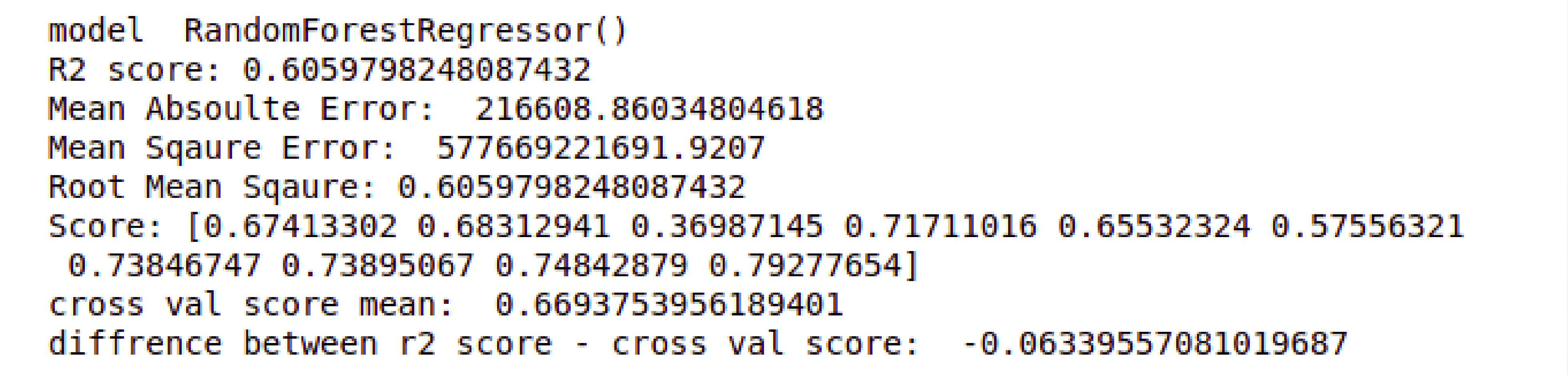
```
model  DecisionTreeRegressor()
R2 score: 0.5277482793915418
Mean Absoulte Error:  247292.9650055371
Mean Sqaure Error:  692363744455.8242
Root Mean Sqaure: 0.5277482793915418
Score: [ 0.02182404  0.61943599  0.3011355   0.43788587  0.35046998 -0.12045959
 -0.82858451  0.48985287  0.61436108  0.21418182]
cross val score mean:  0.21001030487244882
diffrence between r2 score - cross val score:  0.31773797451909297
```

K Neighbors Regressor



```
model  KNeighborsRegressor()
R2 score: 0.40788956273212906
Mean Absoulte Error:  345316.2728239203
Mean Sqaure Error:  868087466044.5146
Root Mean Sqaure: 0.40788956273212906
Score: [0.49476313 0.50380475 0.29380468 0.58503587 0.56844215 0.31260534
 0.46433207 0.45129611 0.63297159 0.41603055]
cross val score mean:  0.4723086232526013
diffrence between r2 score - cross val score:  -0.06441906052047225
```

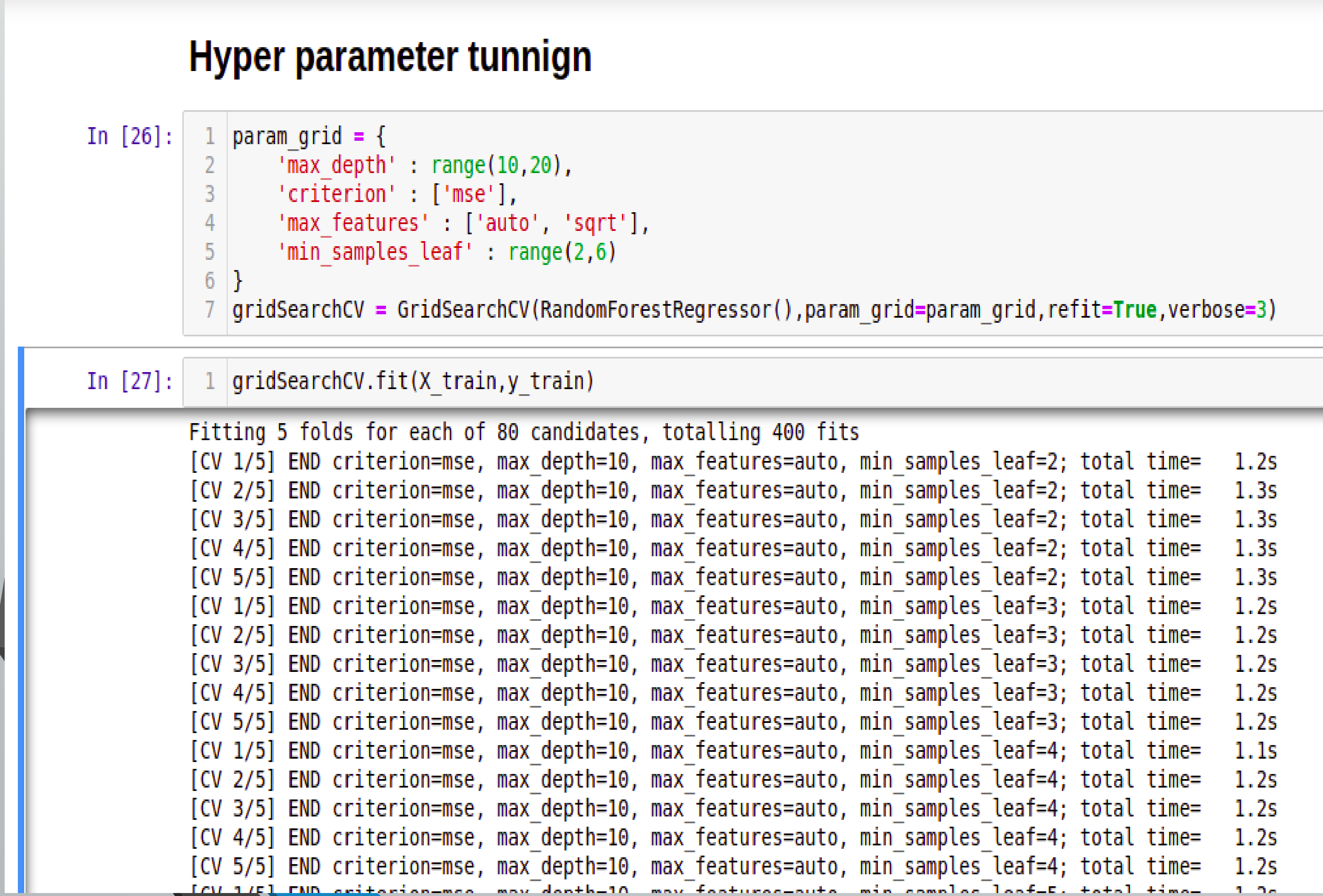
Random Forest Regressor



```
model  RandomForestRegressor()
R2 score: 0.6059798248087432
Mean Absoulte Error:  216608.86034804618
Mean Sqaure Error:  577669221691.9207
Root Mean Sqaure: 0.6059798248087432
Score: [0.67413302 0.68312941 0.36987145 0.71711016 0.65532324 0.57556321
 0.73846747 0.73895067 0.74842879 0.79277654]
cross val score mean:  0.6693753956189401
diffrence between r2 score - cross val score:  -0.06339557081019687
```

RESULT

From the details on the above solutions it is clearly understandable that we are getting best result with the help of Random Forest Regressor so we save this model with the help of joblib Library.

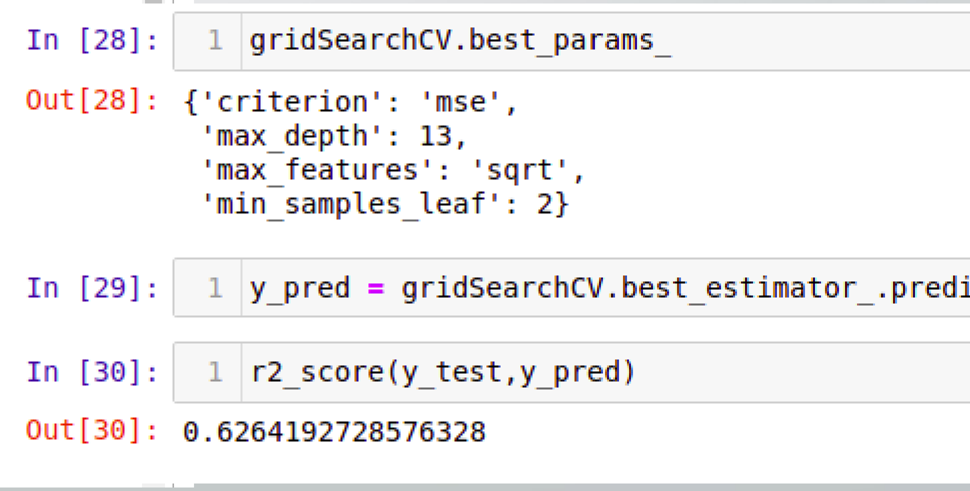


Hyper parameter tunnign

In [26]:
```
param_grid = {
    'max_depth' : range(10,20),
    'criterion' : ['mse'],
    'max_features' : ['auto', 'sqrt'],
    'min_samples_leaf' : range(2,6)
}
gridSearchCV = GridSearchCV(RandomForestRegressor(),param_grid=param_grid,refit=True,verbose=3)
```

In [27]:
```
gridSearchCV.fit(X_train,y_train)
```

```
Fitting 5 folds for each of 80 candidates, totalling 400 fits
[CV 1/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=2; total time=   1.2s
[CV 2/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=2; total time=   1.3s
[CV 3/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=2; total time=   1.3s
[CV 4/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=2; total time=   1.3s
[CV 5/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=2; total time=   1.3s
[CV 1/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=3; total time=   1.2s
[CV 2/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=3; total time=   1.2s
[CV 3/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=3; total time=   1.2s
[CV 4/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=3; total time=   1.2s
[CV 5/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=3; total time=   1.2s
[CV 1/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=4; total time=   1.1s
[CV 2/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=4; total time=   1.2s
[CV 3/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=4; total time=   1.2s
[CV 4/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=4; total time=   1.2s
[CV 5/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=4; total time=   1.2s
[CV 1/5] END criterion=mse, max_depth=10, max_features=auto, min_samples_leaf=5; total time=   1.2s
```



In [28]:
```
gridSearchCV.best_params_
```

Out[28]:
```
{'criterion': 'mse',
 'max_depth': 13,
 'max_features': 'sqrt',
 'min_samples_leaf': 2}
```

In [29]:
```
y_pred = gridSearchCV.best_estimator_.predi
```

In [30]:
```
r2_score(y_test,y_pred)
```

Out[30]:
```
0.6264192728576328
```

CONCLUSION

- ✓ From this dataset I get to know that each feature plays a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithm

- ✓ The power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important steps to remove missing value or null value fill it by mean median or by mode or by 0.

- ✓ Various algorithms I used in this dataset and to get out best result and save that model. The best algorithm is Random Forest Regressor.

FUTURE WORK

- ✓ Limitations of this project is we have less number of features. If we get interior column, where we will get feature like, A/C, air bag etc. More the number of features, more accuracy we'll get.

- ✓ In future, if someone do the proper and detail study of this dataset's each column than the accuracy will be so high.

Thank You