House Prices - Advanced Regression Techniques

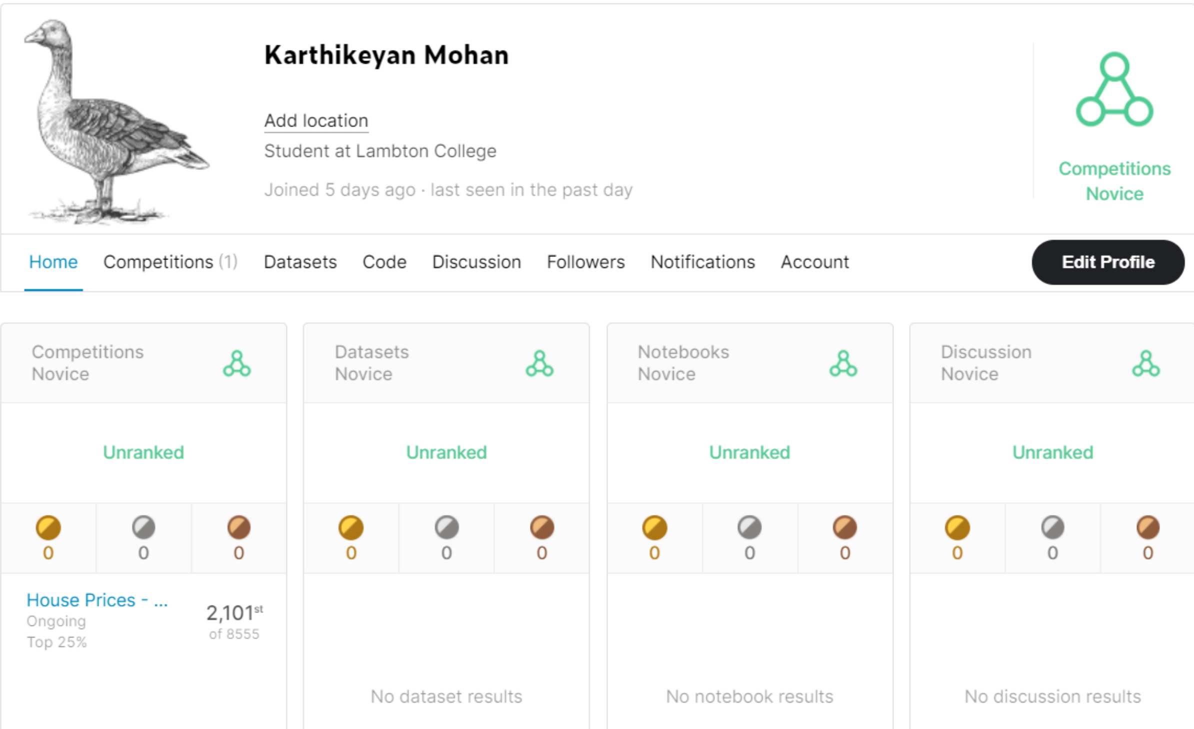
Kaggle Competition

Presented to

Prof. Moez Ali
Data Science
Lambton College

April 14, 2021

Presented by Karthikeyan Mohan


Karthikeyan Mohan
Add location
Student at Lambton College
Joined 5 days ago · last seen in the past day
Competitions
Novice
Home
Competitions (1)
Datasets
Code
Discussion
Followers
Notifications
Account
Edit Profile
Competitions Novice
Unranked
0 0 0
House Prices - ...
Ongoing
Top 25%
2,101st
of 8555
Datasets Novice
Unranked
0 0 0
No dataset results
Notebooks Novice
Unranked
0 0 0
No notebook results
Discussion Novice
Unranked
0 0 0
No discussion results

Contents

A. Project Background:

In this project, we have to build the machine learning model to predict the sale price of the house based on 80 attribures present in the dataset. The key prerequisites in this project are a dataset containing house related attributes, python libraries, various machine learning algorithms, visualization packages and the pycaret library. The ultimate goal of the project is to predict the saleprice of the test data given in the Kaggle competition and submit the results to achieve the best Kaggle score.

B. Model Design Approach :

Before starting working on the project, brainstormed the dataset descriptions and listed all the steps required to get the required end results. Below are the design steps carried out :

- Understand the each attribute in the dataset and find their datatypes and the values present in each. Identify the target variables to predict.
- Used python visualisation libraries to visualize the data and find the pattern and relations between each attribute.
- Split the dataset for test and train.
- Based on the exploratory data analysis, perform the data cleansing and remove the noise in the data.
- Perform the transformation and scaling technique if required.
- Assign the input and output into the separate variables for the model input.

- Build the model and identified the best performing algorithms using pycaret by evaluate the model.
- Predict the saleprice for the test data provided in the Kaggle and submit the result.
- Make changes in preprocessing and fine tune the model until we get the expected score in Kaggle.

C. Exploratory data analysis :

The analysis carried out in this step can be simply segregrated into below caetgories and the key observations are summaried at the end of this section.

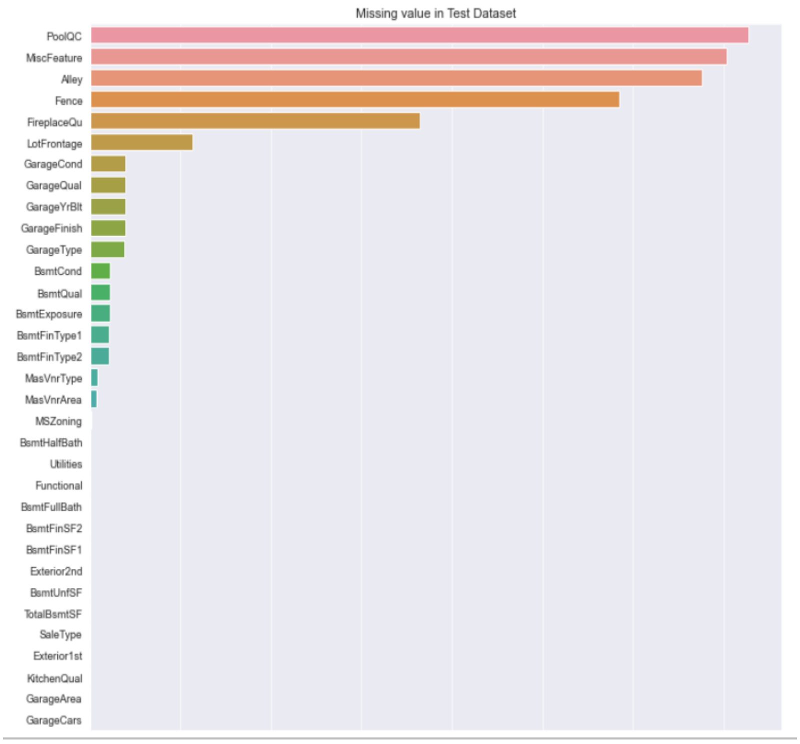
1. **Basic Statistics -** Executed few pandas commands to find out the records counts, datatypes, mean,mode,standard deviation of data

```
data.shape,test_data.shape
((1460, 81), (1459, 80))
data.info()
test_data.info()
data.describe().transpose()
test_data.describe().transpose()
```

2. **Finding Missing values – Listing out the missing values in the training and test dataset.**

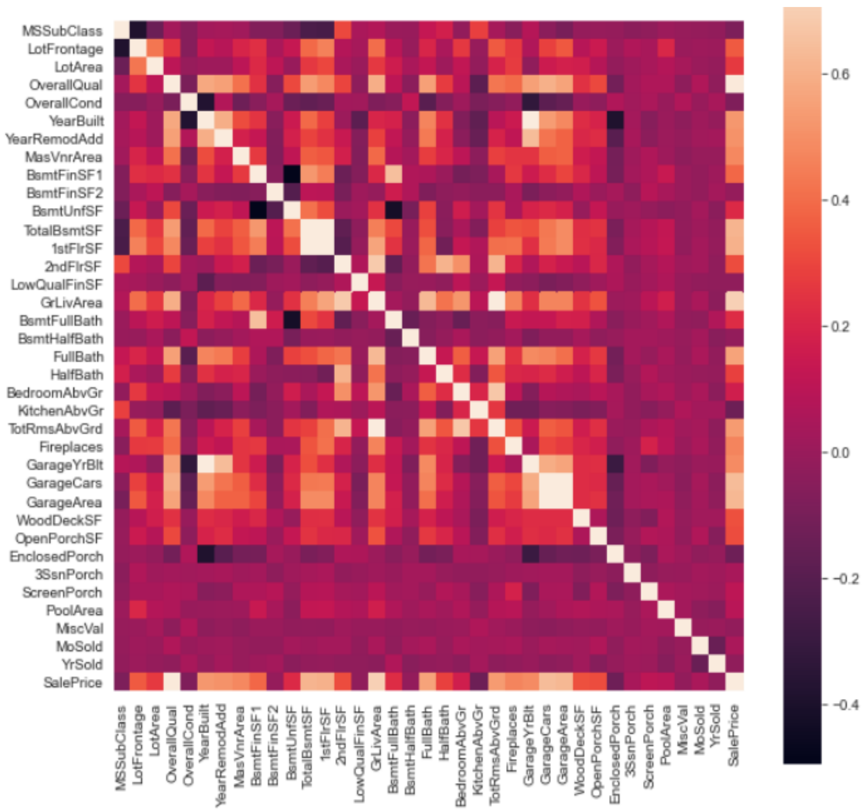
```
Missing values in Train
PoolQC          1453
MiscFeature     1406
Alley           1369
Fence           1179
FireplaceQu      690
LotFrontage      259
GarageCond        81
GarageType        81
GarageYrBlt       81
GarageFinish      81
GarageQual        81
BsmtExposure      38
BsmtFinType2      38
BsmtFinType1      37
BsmtCond          37
BsmtQual          37
MasVnrArea         8
MasVnrType         8
Electrical         1
dtype: int64
```

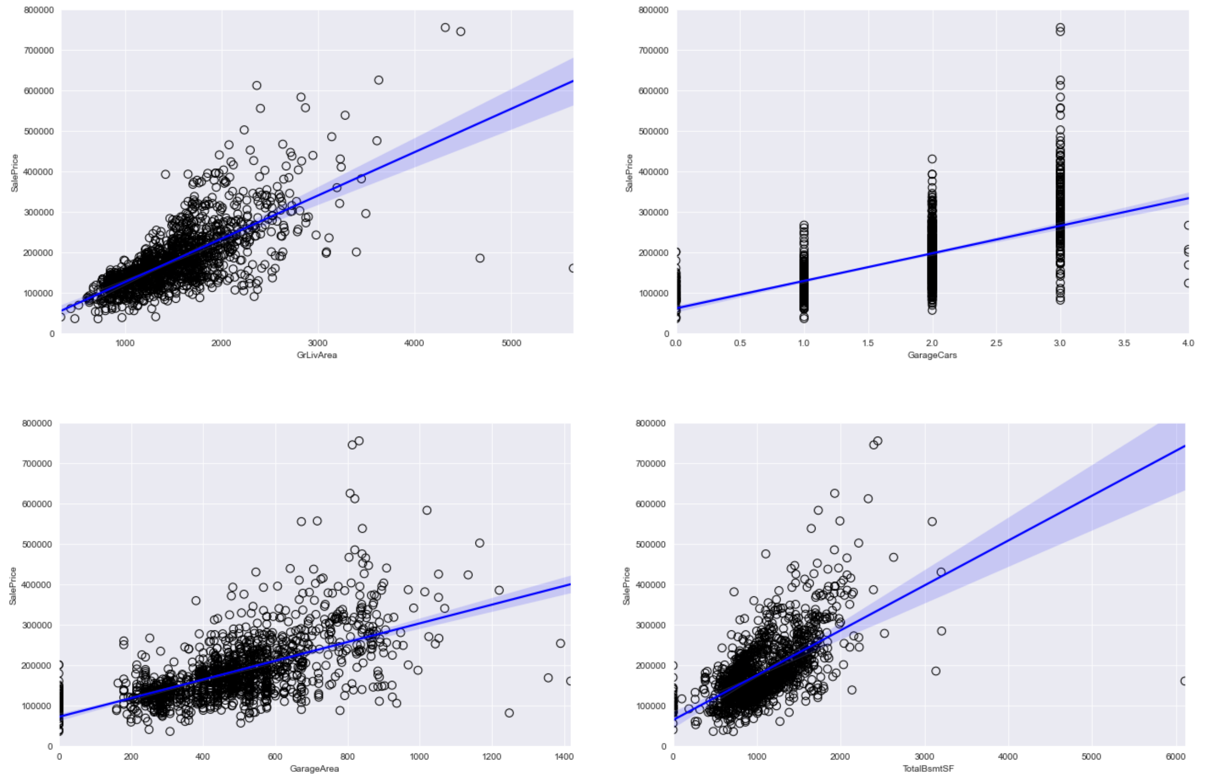
```
Missing values in Test data
PoolQC          1456
MiscFeature     1408
Alley           1352
Fence           1169
FireplaceQu      730
LotFrontage      227
GarageCond        78
GarageQual        78
GarageYrBlt       78
GarageFinish      78
GarageType        76
BsmtCond          45
BsmtQual          44
BsmtExposure      44
BsmtFinType1      42
BsmtFinType2      42
MasVnrType        16
MasVnrArea        15
MSZoning           4
BsmtHalfBath       2
Utilities          2
Functional         2
BsmtFullBath       2
BsmtFinSF2         1
BsmtFinSF1         1
Exterior2nd        1
BsmtUnfSF          1
TotalBsmtSF        1
SaleType           1
Exterior1st        1
KitchenQual        1
GarageArea         1
GarageCars         1
dtype: int64
```



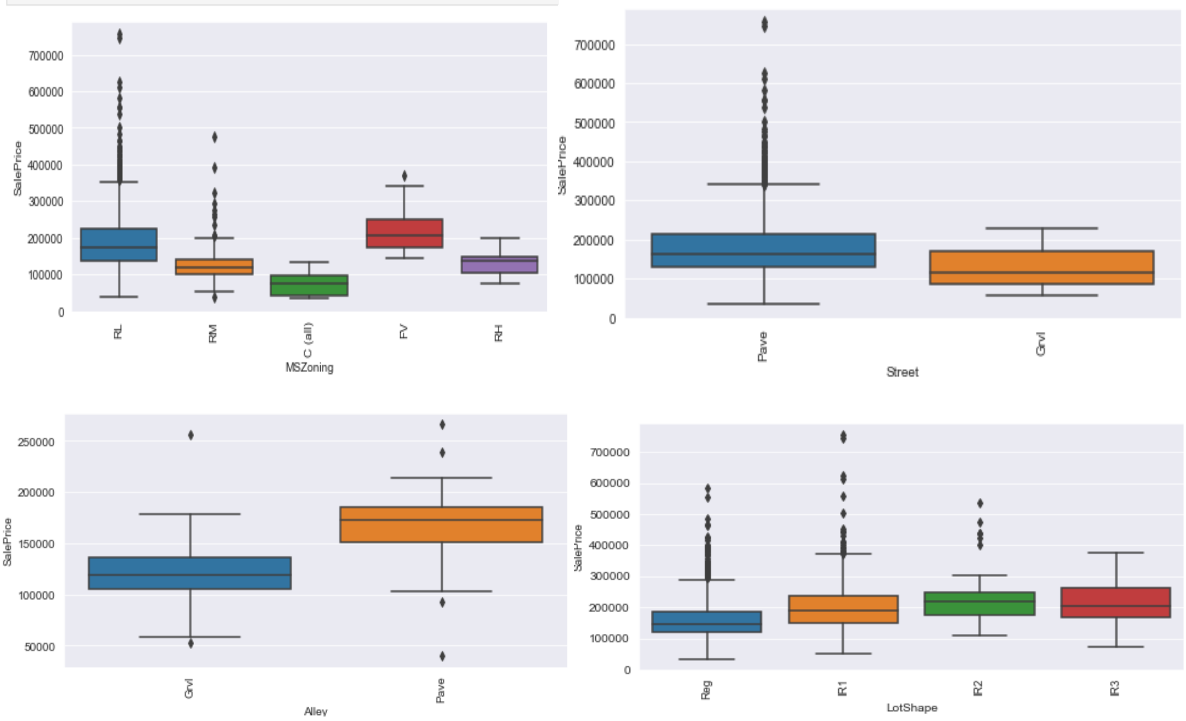
3. Numerical data analysis:

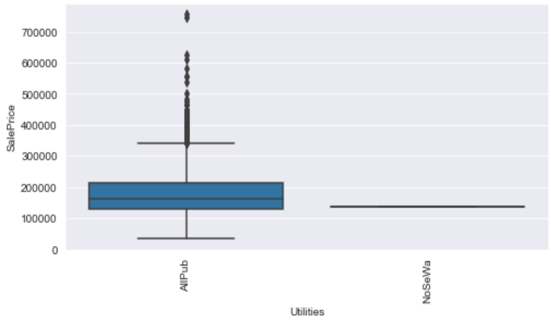
```
OverallQual      0.790982
GrLivArea        0.708624
GarageCars       0.640409
GarageArea       0.623431
TotalBsmtSF      0.613581
1stFlrSF         0.605852
FullBath         0.560664
TotRmsAbvGrd     0.533723
YearBuilt        0.522897
YearRemodAdd     0.507101
GarageYrBlt      0.486362
MasVnrArea       0.477493
Fireplaces       0.466929
BsmtFinSF1       0.386420
LotFrontage      0.351799
WoodDeckSF       0.324413
2ndFlrSF         0.319334
OpenPorchSF      0.315856
HalfBath         0.284108
LotArea          0.263843
BsmtFullBath     0.227122
BsmtUnfSF        0.214479
BedroomAbvGr     0.168213
ScreenPorch      0.111447
PoolArea         0.092404
MoSold           0.046432
3SsnPorch        0.044584
BsmtFinSF2      -0.011378
BsmtHalfBath    -0.016844
MiscVal         -0.021190
LowQualFinSF    -0.025606
YrSold          -0.028923
OverallCond     -0.077856
MSSubClass      -0.084284
EnclosedPorch   -0.128578
KitchenAbvGr    -0.135907
```

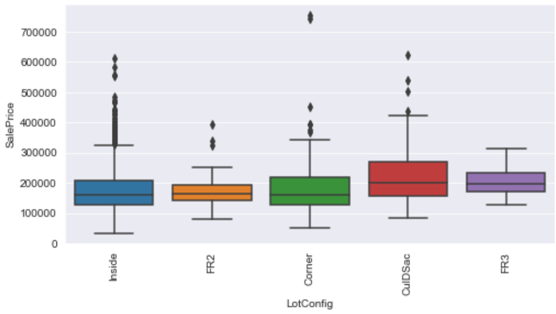


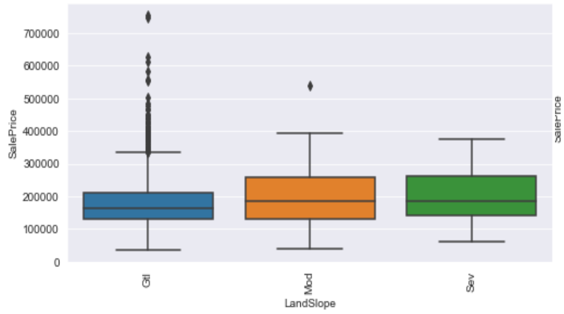


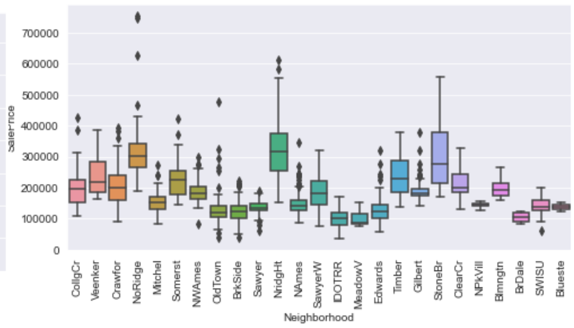
4. Categorical data analysis

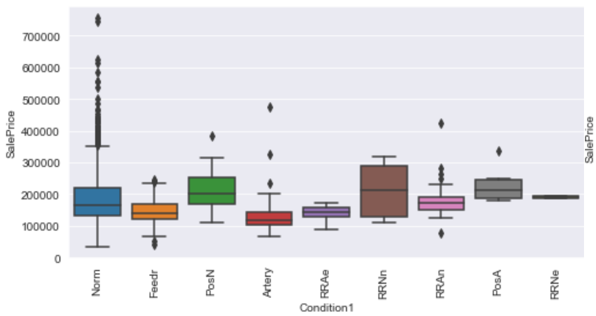


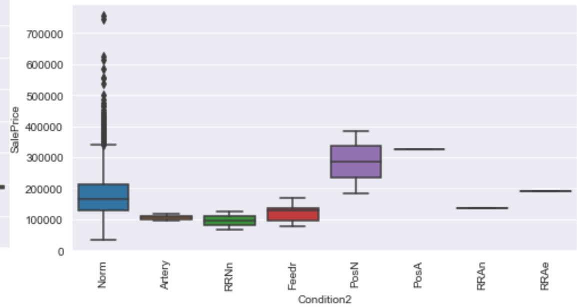

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
AllPub
NoSeWa
Utilities

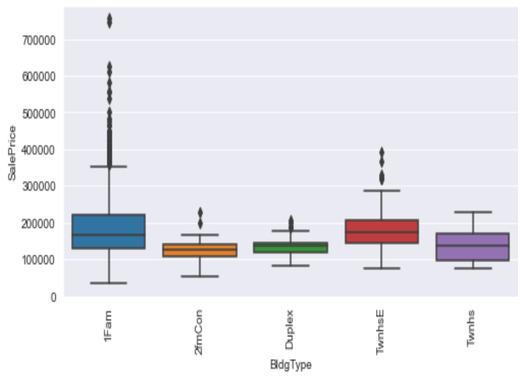

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Inside
FR2
Corner
CulDSac
FR3
LotConfig

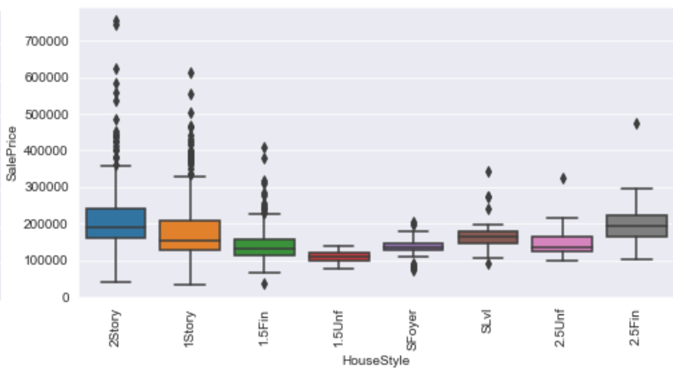

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gtl
Mod
Sev
LandSlope

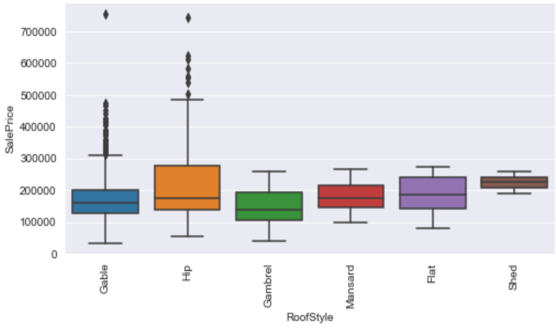

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
CollgCr
Veenker
Crawfor
NoRidge
Mitchel
Somerst
NWAmes
OldTown
BrkSide
Sawyer
NridgHt
NAmes
SawyerW
IDOTRR
MeadowV
Edwards
Timber
Gilbert
StoneBr
ClearCr
NPkVill
Blmngtn
BrDale
SWISU
Blueste
Neighborhood

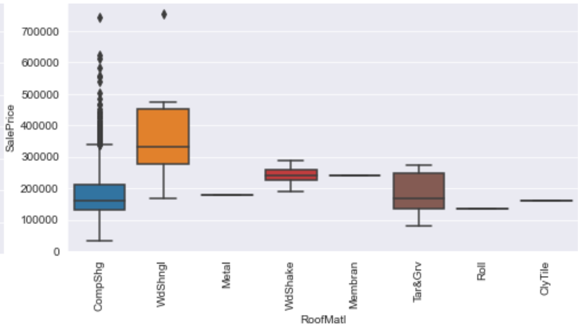

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Norm
Feedr
PosN
Artery
RRAe
RRNn
RRAn
PosA
RRNe
Condition1

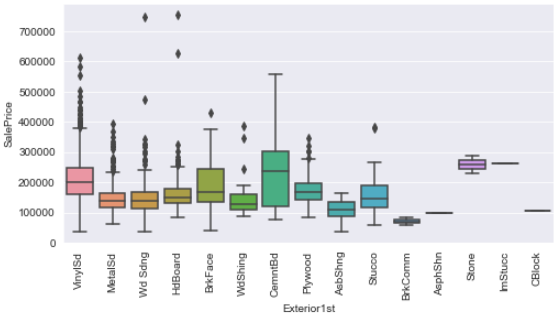

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Norm
Artery
RRNn
Feedr
PosN
PosA
RRAn
RRAe
Condition2

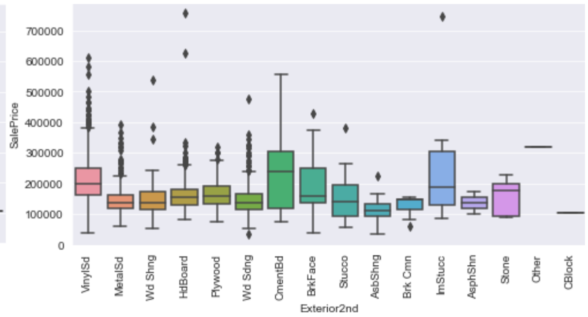

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1Fam
2fmCon
Duplex
TwnhsE
Twnhs
BldgType

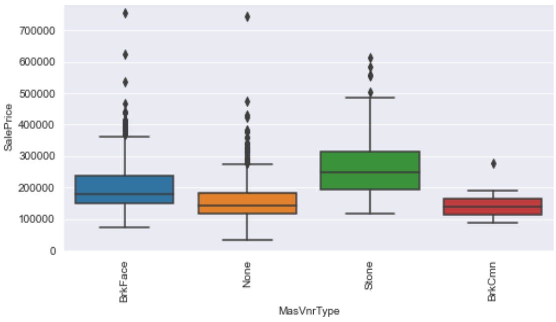

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
2Story
1Story
1.5Fin
1.5Unf
SFoyer
SLvl
2.5Unf
2.5Fin
HouseStyle

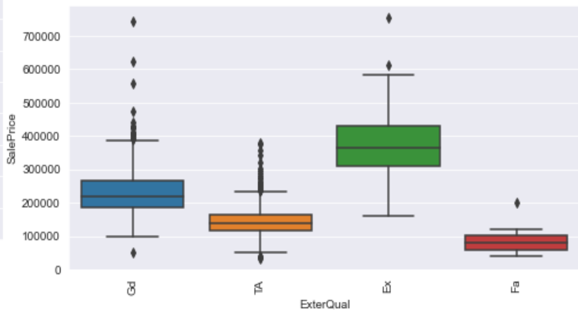

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gable
Hip
Gambrel
Mansard
Flat
Shed
RoofStyle

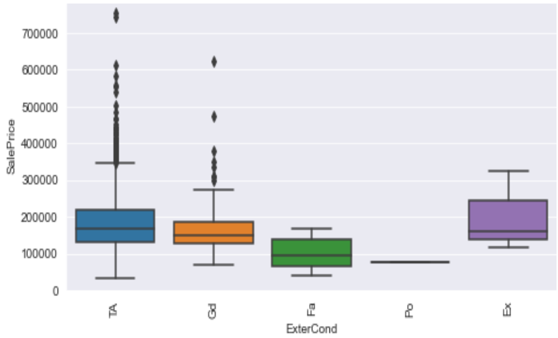

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
CompShg
WdShngl
Metal
WdShake
Membran
Tar&Grv
Roll
ClyTile
RoofMatl

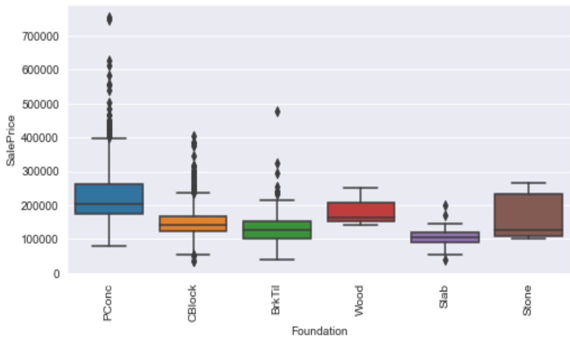

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
VinylSd
MetalSd
Wd Sdng
HdBoard
BrkFace
WdShing
CemntBd
Plywood
AsbShng
Stucco
BrkComm
AsphShn
Stone
ImStucc
CBlock
Exterior1st

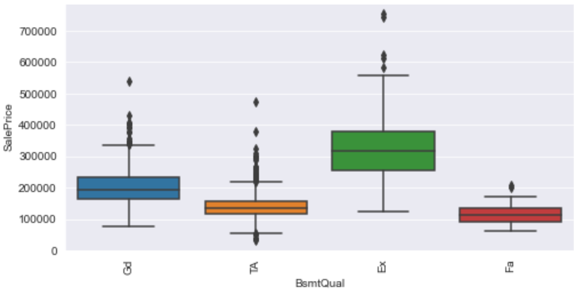

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
VinylSd
MetalSd
Wd Shng
HdBoard
Plywood
Wd Sdng
CmentBd
BrkFace
Stucco
AsbShng
Brk Cmn
ImStucc
AsphShn
Stone
Other
CBlock
Exterior2nd

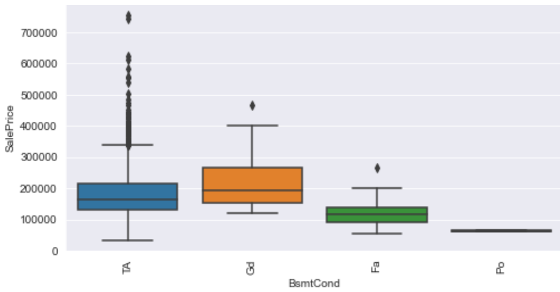

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
BrkFace
None
Stone
BrkCmn
MasVnrType

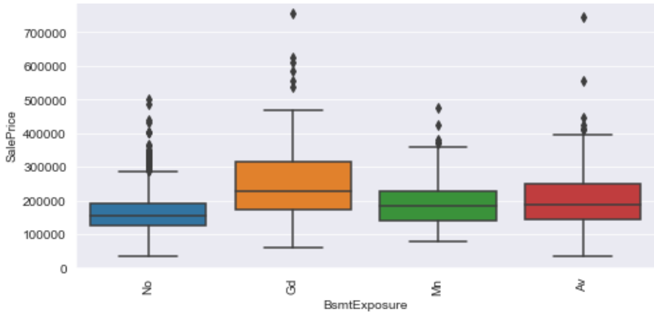

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gd
TA
Ex
Fa
ExterQual

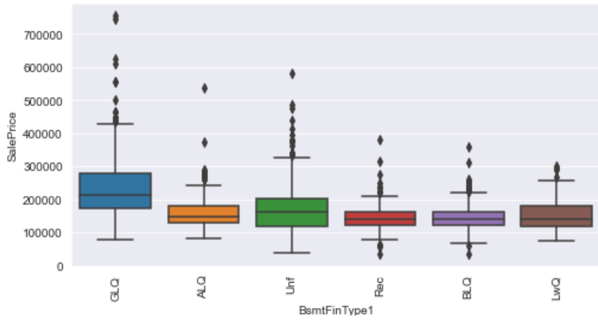

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Gd
Fa
Po
Ex
ExterCond

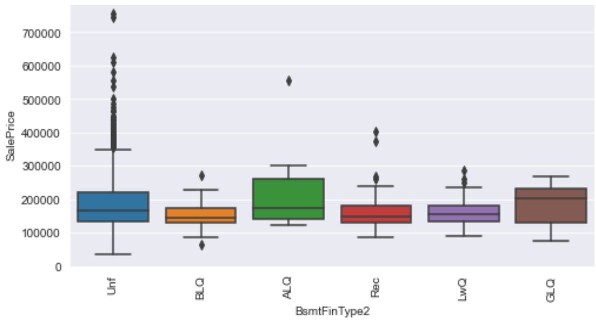

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
PConc
CBlock
BrkTil
Wood
Slab
Stone
Foundation

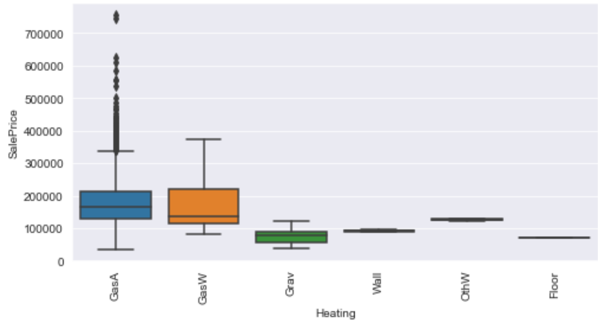

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gd
TA
Ex
Fa
BsmtQual

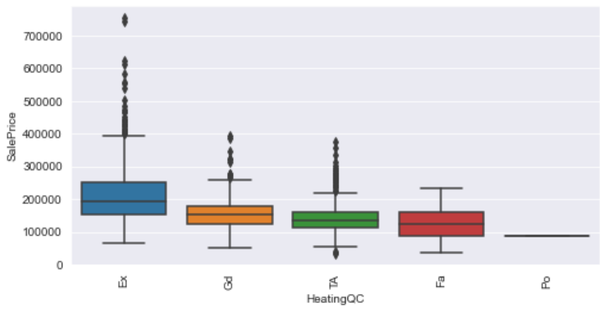

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Gd
Fa
Po
BsmtCond

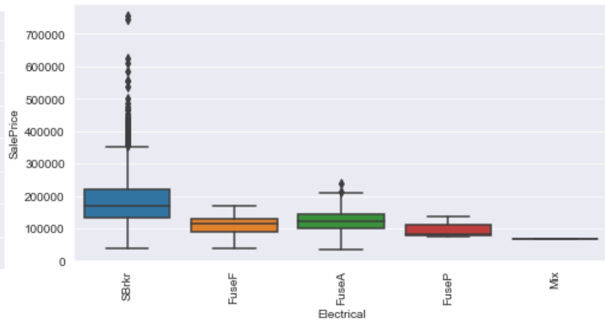

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
No
Gd
Mn
Av
BsmtExposure

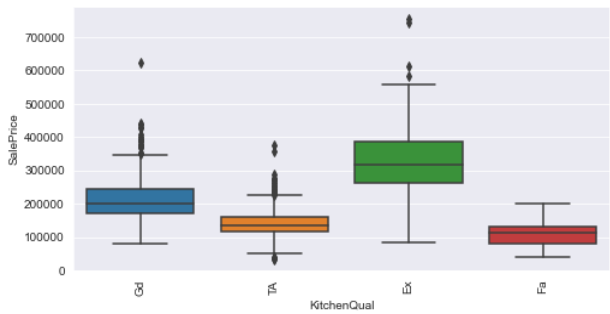

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
GLQ
ALQ
Unf
Rec
BLQ
LwQ
BsmtFinType1

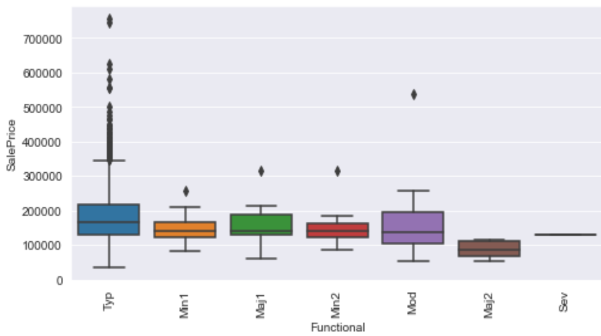

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Unf
BLQ
ALQ
Rec
LwQ
GLQ
BsmtFinType2

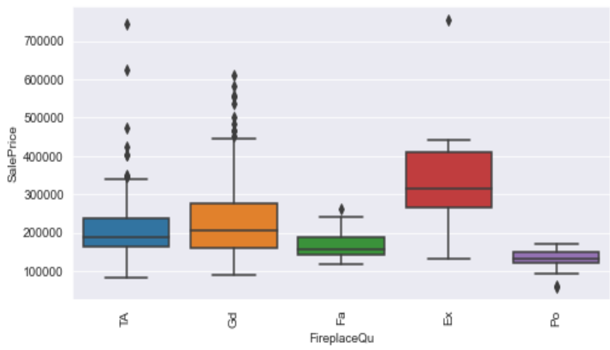

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
GasA
GasW
Grav
Wall
OthW
Floor
Heating

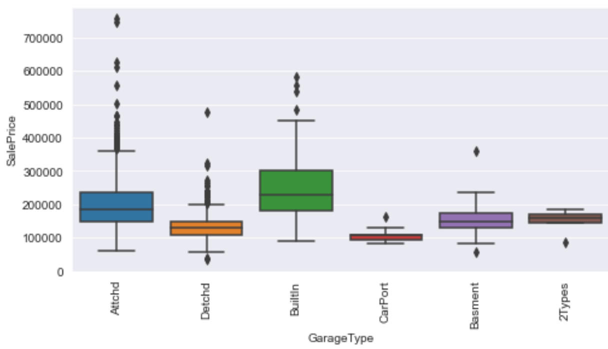

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Ex
Gd
TA
Fa
Po
HeatingQC

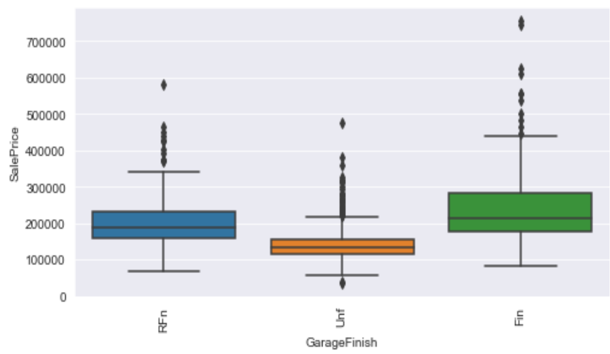

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
SBrkr
FuseF
FuseA
FuseP
Mix
Electrical

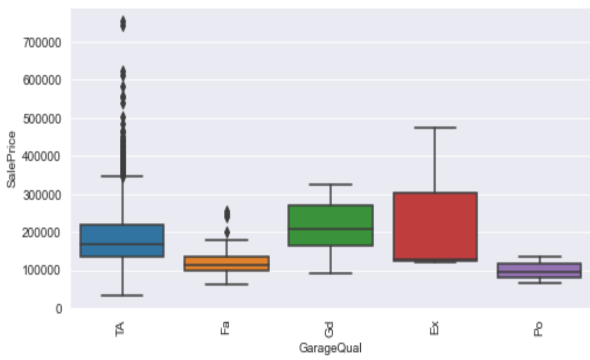

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gd
TA
Ex
Fa
KitchenQual

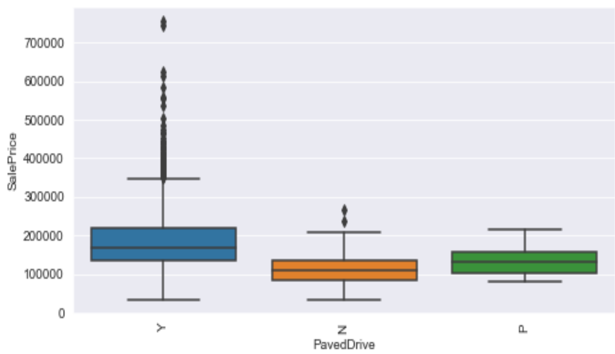

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Typ
Min1
Maj1
Min2
Mod
Maj2
Sev
Functional

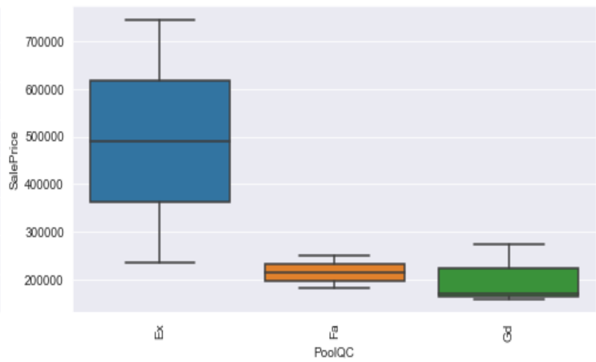

SalePrice
100000
200000
300000
400000
500000
600000
700000
TA
Gd
Fa
Ex
Po
FireplaceQu

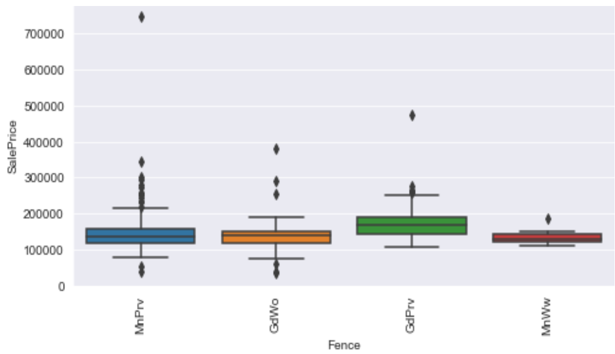

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Attchd
Detchd
BuiltIn
CarPort
Basment
2Types
GarageType

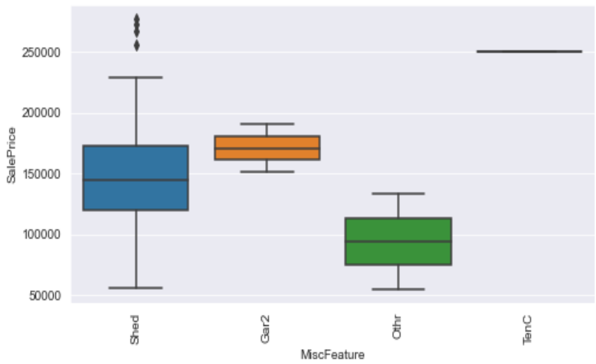

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
RFn
Unf
Fin
GarageFinish

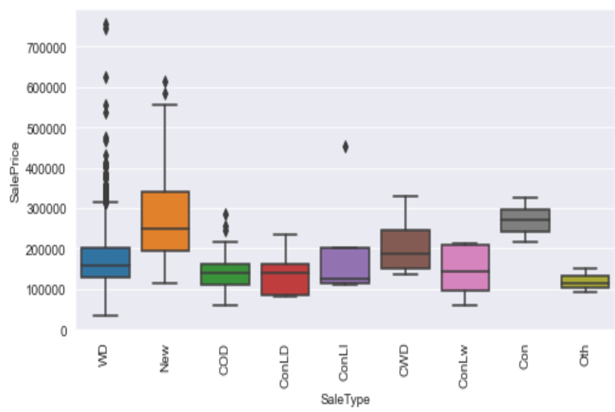

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
TA
Fa
Gd
Ex
Po
GarageQual

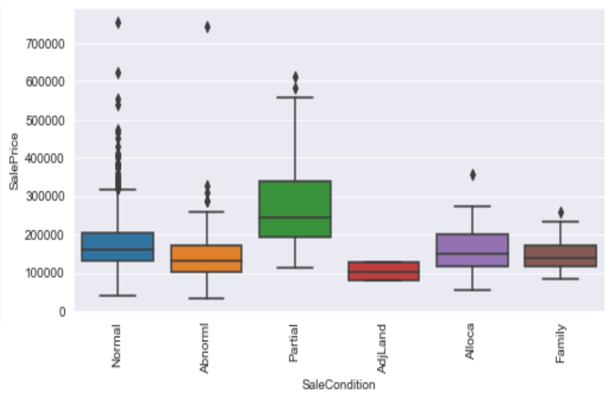

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Y
N
P
PavedDrive


SalePrice
700000
600000
500000
400000
300000
200000
Ex
Fa
Gd
PoolQC

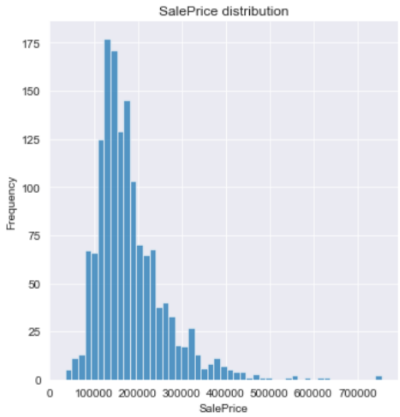

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
MnPrv
GdWo
GdPrv
MnWw
Fence


SalePrice
250000
200000
150000
100000
50000
Shed
Gar2
Othr
TenC
MiscFeature


SalePrice
700000
600000
500000
400000
300000
200000
100000
0
WD
New
COD
ConLD
ConLI
CWD
ConLw
Con
Oth
SaleType


SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Normal
Abnorml
Partial
AdjLand
Alloca
Family
SaleCondition

5. Target value analysis



Summary of EDA :

1. 1460 Records and 81 fields present in the training dataset. 1459 Records and 80 fields present in the training dataset.

2. First field 'ID' is the sequence which doesn't add any value to our prediction and hence can be dropped from the dataset.

3. Missing values present in 19 fields in train set and 25 fields in test set. By perform EDA on the data in those fields, we could identify to use mode or mean or perform imperative iteration technique to impute the data.

4. As we expected, saleprice is mainly determined by the overall quality of the house. OverallQual variable is highly correlated with the saleprice.

5. Independent fields like GrLivArea, GarageCars and GarageArea also have good correlation with target field Sale Price. GrLivArea & TotRmsAbvGrd, YearBuilt & GarageYrBuilt, GarageCars & GarageArea, 1stFlrSF & TotalBsmtSF are correlated among themselves.

6. There are some ordinal categorical data such as PoolQC, ExterQual and ExterCond,BSMTQual can be transformed to the numerical data. This will improve the accuracy of the model.

7. SalePrice is not normally distributed equally, we have to any of the scaling method to normalise the data and distribute equally.

8. We created a function for each preprocessing steps and reused it for test data to preprocess.

D. Data Preprocessing :

1. Handling Missing value

Below are the logic used imputing the Null values :

Variable	Impute Logic
PoolQC	Filled NULL with 0
MiscFeature	Filled NULL with NA
Alley	Filled NULL with NA
Fence	Filled NULL with NA
FireplaceQu	Filled NULL with 0
LotFrontage	Took mean of LotArea and Lotfrontage and then divided Lotarea to the calculated mean
GarageQual	Filled NULL with 0
GarageYrBlt	Filled NULL with 0
GarageType	Filled NULL with NA
GarageCond	Filled NULL with 0
GarageFinish	Filled NULL with 0
BsmtFinType2	Filled NULL with 0
BsmtFinType1	Filled NULL with 0
BsmtExposure	Filled NULL with 0
BsmtQual	Filled NULL with 0
BsmtCond	Filled NULL with 0
MasVnrType	Filled NULL with 0
MasVnrArea	Filled NULL with 0
Electrical	Filled NULL with SBrkr
BsmtHalfBath	Filled NULL with 0
BsmtFullBath	Filled NULL with 0
TotalBsmtSF	Filled NULL with Mean
GarageArea	Filled NULL with Mean
BsmtUnfSF	Filled NULL with Mean
GarageCars	Filled NULL with 2
BsmtFinSF2	Filled NULL with 0
BsmtFinSF1	Filled NULL with 0
MSZoning	Filled NULL with Mode
Functional	Filled NULL with Typ
Utilities	Filled NULL with AllPub
Saletype	Filled NULL with Mode
Exterior1st	Filled NULL with Mode
Exterior2nd	Filled NULL with Mode
KitchenQual	Filled NULL with Mode

2. Tranform ordinal data

There are multiple ordinal categorical data in our dataset, we picked only three variables based on our EDA. This helped us to increase the accuracy and score of our Kaggle result.

- PoolQC
- ExterQual
- ExterCond

Below is the logic used to replace the data into ordindal data for all 3 fields.

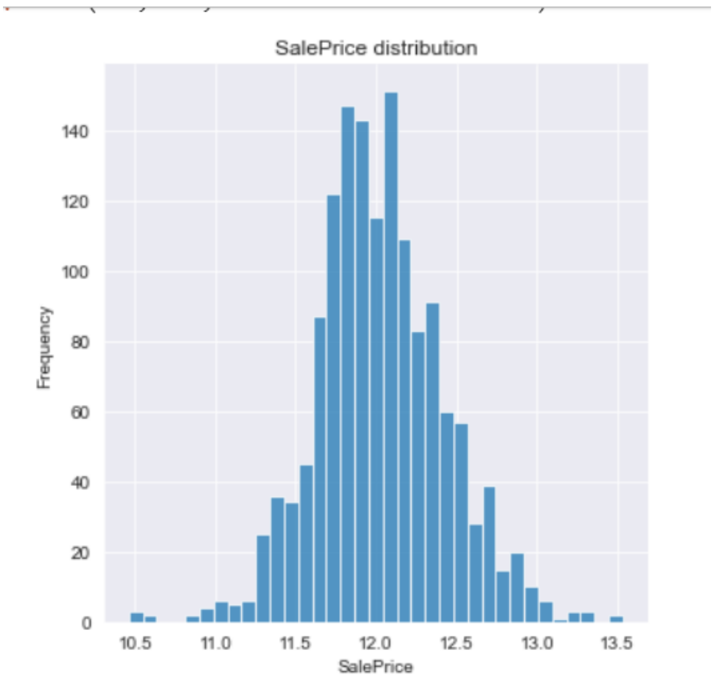
dataset_name['PoolQC'] = dataset_name['PoolQC'] .replace('Ex',5, regex=True)
dataset_name['PoolQC'] = dataset_name['PoolQC'] .replace('Gd',4, regex=True)
dataset_name['PoolQC'] = dataset_name['PoolQC'] .replace('TA',3, regex=True)
dataset_name['PoolQC'] = dataset_name['PoolQC'] .replace('Fa',2, regex=True)
dataset_name['PoolQC'] = dataset_name['PoolQC'] .replace('Po',1, regex=True)

3. Normalise Target value

We used log transformation method to distribute the data in equal manner.

data["SalePrice"] = np.log1p(data["SalePrice"])

Saleprice aftrer distribution



E. Data Modeling:

1. Identifying the best algorithm:

Thanks to Pycaret !! we used this low code machine learning libray to find the best model for our dataset in a minute of setup and execution. This helped us save our effort and time. Below is the list of model with its accuracy.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.0836	0.0159	0.1250	0.8942	0.0097	0.0070	2.7190
br	Bayesian Ridge	0.0843	0.0186	0.1321	0.8797	0.0102	0.0071	0.1170
gbr	Gradient Boosting Regressor	0.0926	0.0186	0.1352	0.8769	0.0105	0.0077	0.1970
omp	Orthogonal Matching Pursuit	0.0912	0.0193	0.1363	0.8727	0.0105	0.0076	0.0250
lightgbm	Light Gradient Boosting Machine	0.0957	0.0200	0.1403	0.8673	0.0109	0.0080	0.1180
ridge	Ridge Regression	0.0895	0.0205	0.1391	0.8660	0.0107	0.0075	0.0290
rf	Random Forest Regressor	0.0995	0.0217	0.1459	0.8573	0.0113	0.0083	0.3630
xgboost	Extreme Gradient Boosting	0.1005	0.0226	0.1491	0.8497	0.0116	0.0084	0.4990
lr	Linear Regression	0.1024	0.0257	0.1566	0.8305	0.0121	0.0086	0.4070
huber	Huber Regressor	0.1122	0.0285	0.1667	0.8122	0.0128	0.0094	0.2430
et	Extra Trees Regressor	0.1102	0.0282	0.1666	0.8110	0.0129	0.0092	0.4510
en	Elastic Net	0.1141	0.0292	0.1689	0.8076	0.0130	0.0096	0.0250
lasso	Lasso Regression	0.1204	0.0316	0.1760	0.7914	0.0135	0.0101	0.0280
ada	AdaBoost Regressor	0.1363	0.0325	0.1793	0.7846	0.0138	0.0114	0.1440
dt	Decision Tree Regressor	0.1472	0.0435	0.2079	0.7071	0.0161	0.0123	0.0270
knn	K Neighbors Regressor	0.1628	0.0504	0.2240	0.6625	0.0172	0.0136	0.0360
par	Passive Aggressive Regressor	0.1987	0.0893	0.2738	0.3689	0.0206	0.0165	0.0220
llar	Lasso Least Angle Regression	0.3050	0.1516	0.3882	-0.0091	0.0298	0.0254	0.3880

2. Building the final model

As obvious from the above result, we picked the catboost regression algorithm as our final alogorithm to build our model.

cat_model= CatBoostRegressor()
cat_model.fit(X_train, y,cat_features=cat_feat)

3. Deploying and predicting the test results

At the end, we fit our model into the test data and predicted the saleprice of the test data and submitted in Kaggle. That gave us the best score of

F. Conclusion:

This project made me to involve more practical work on machine learning subject. I could see the bigger picture how the machine learning has been handled.It motivated me to learn and read new concepts and work towards building more machine learning projects.