**House Prices - Advanced Regression Techniques**

Kaggle Competition

Presented to

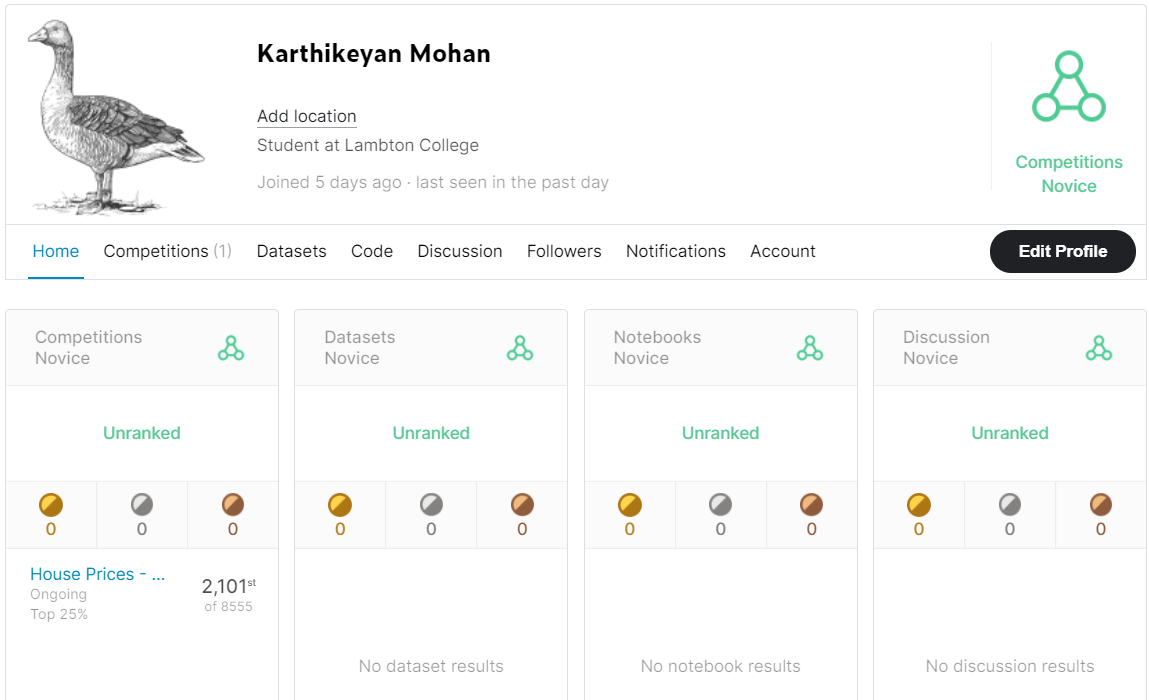
Prof. Moez Ali

Data Science

Lambton College

April 14, 2021

Presented by Karthikeyan Mohan



**Contents**

[A. Project Background: 4](#_Toc69745110)

[B. Model Design Approach : 4](#_Toc69745111)

[C. Exploratory data analysis : 5](#_Toc69745112)

[1. Basic Statistics 5](#_Toc69745113)

[2. Finding Missing values 5](#_Toc69745114)

[3. Categorical data analysis 6](#_Toc69745115)

[4. Numerical data analysis 7](#_Toc69745116)

[5. Target value analysis 12](#_Toc69745117)

[D. Data Preprocessing : 13](#_Toc69745118)

[1. Handling Missing values 13](#_Toc69745119)

[2. Normalise Target value 14](#_Toc69745120)

[3. Tranform ordinal data **Error! Bookmark not defined.**](#_Toc69745121)

[E. Data Modeling: 15](#_Toc69745122)

[F. Future Enhancements: **Error! Bookmark not defined.**](#_Toc69745123)

[G. Conclusion: 15](#_Toc69745124)

[H. Reference : **Error! Bookmark not defined.**](#_Toc69745125)

# 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.

****

# Model Design Approach :

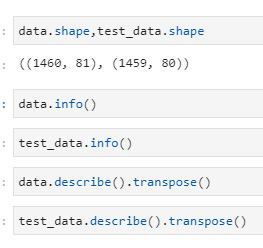
Before starting working on the project, brainstormed the dataset descriptions and listed all the steps required to get the required end results. Below are the design steps carried out :

* 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.

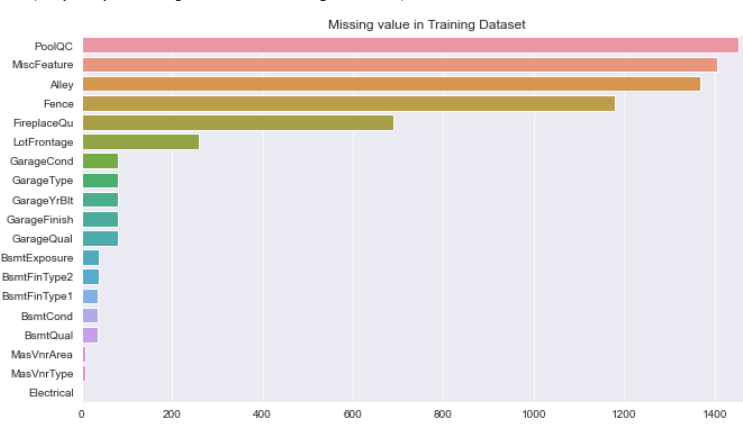
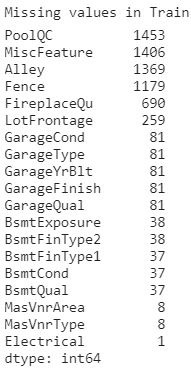
# Exploratory data analysis :

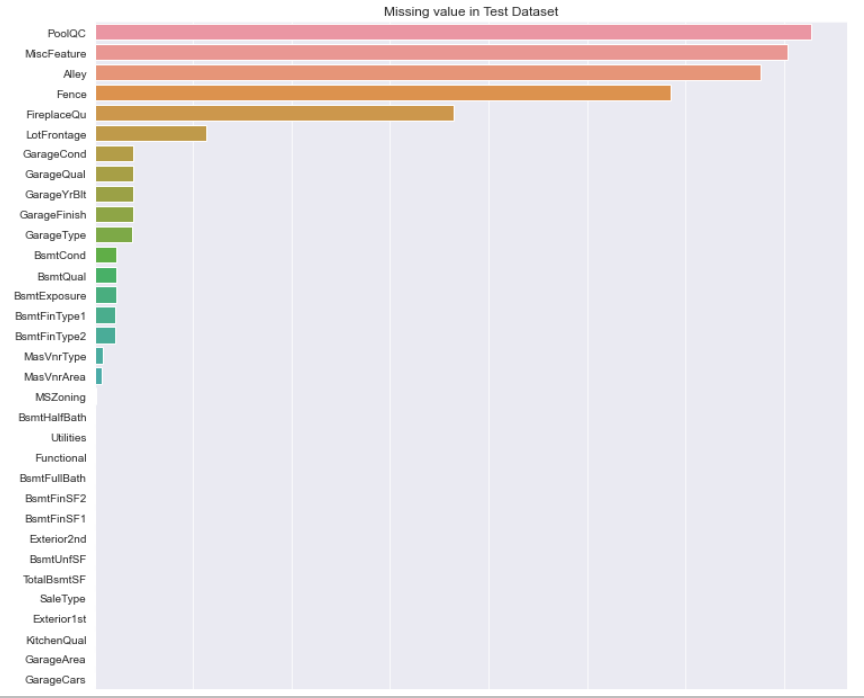
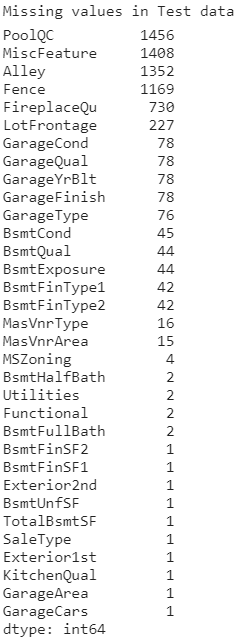
The analysis carried out in this step can be simply segregrated into below caetgories and the key observations are summaried at the end of this section.

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

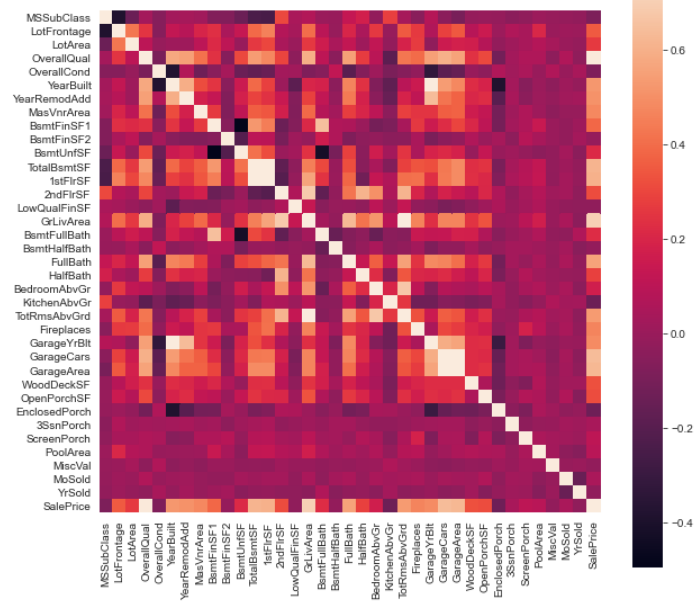
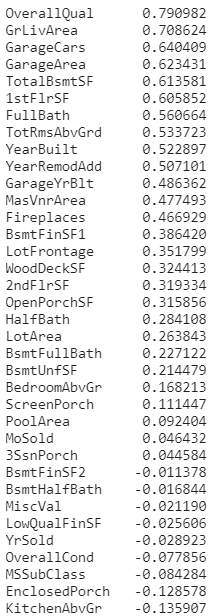


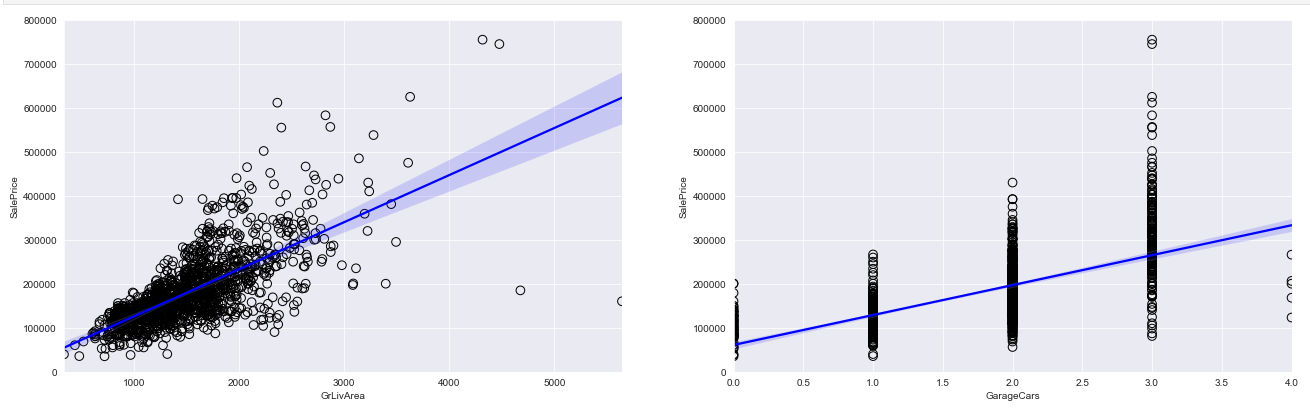
1. **Finding Missing values** – **Listing out the missing values in the training and test dataset.**

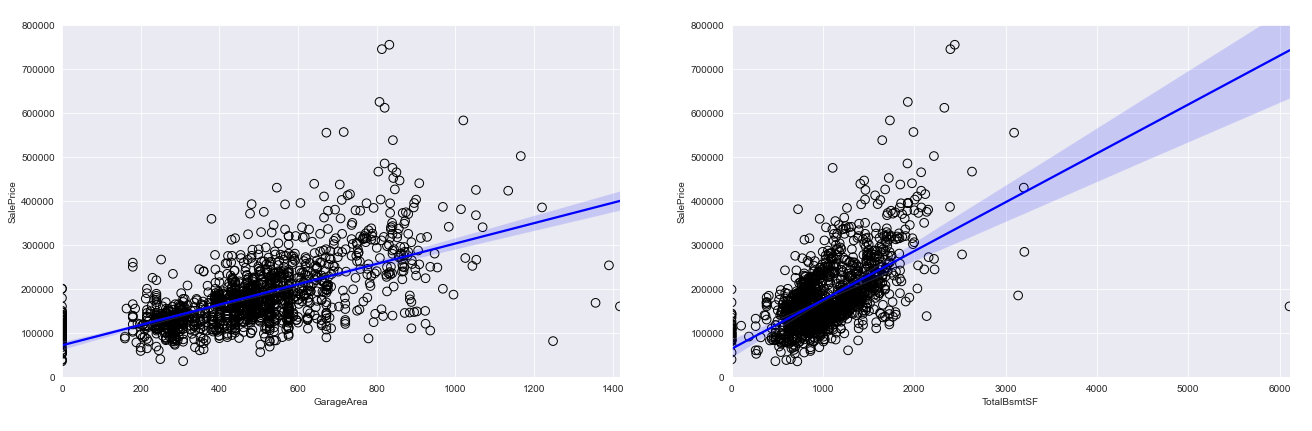




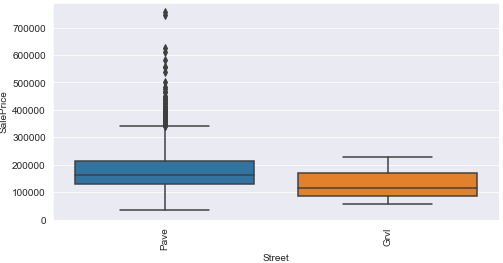
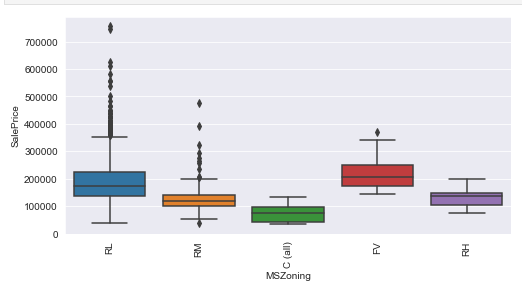
# Numerical data analysis:

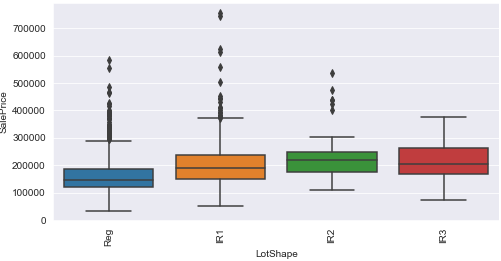
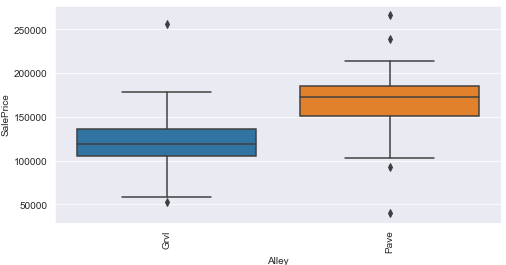


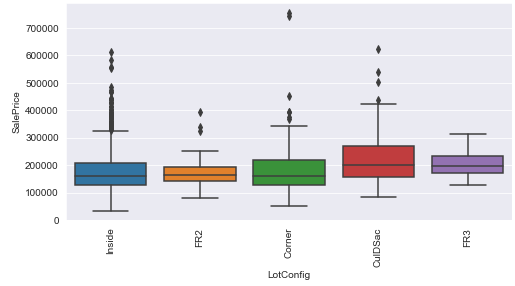
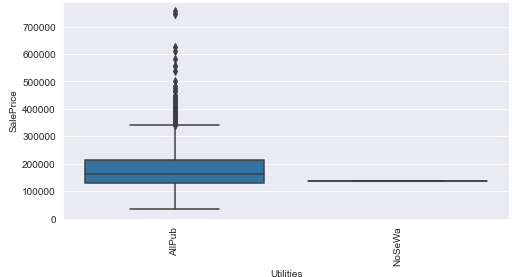


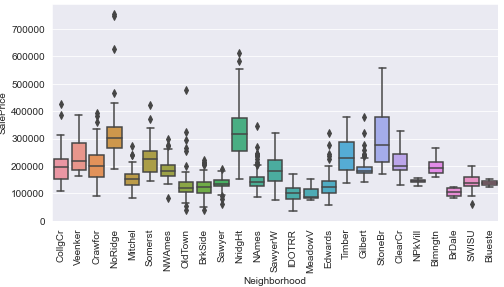
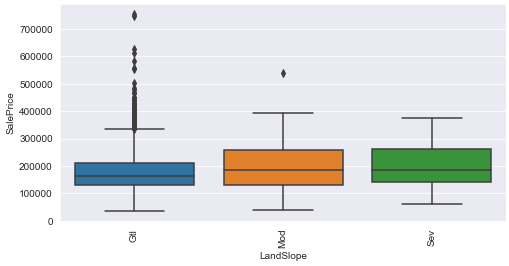


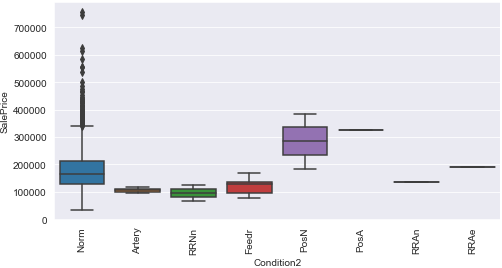
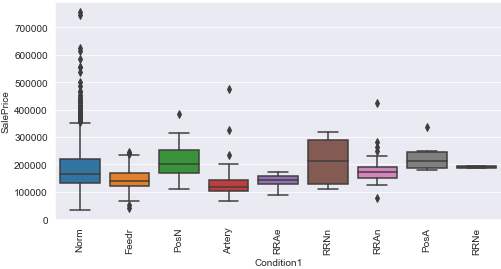
# Categorical data analysis

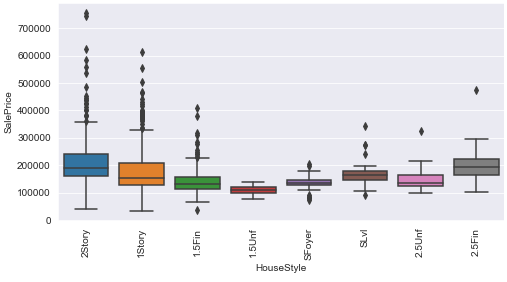
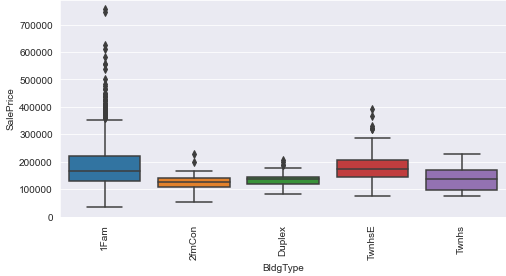


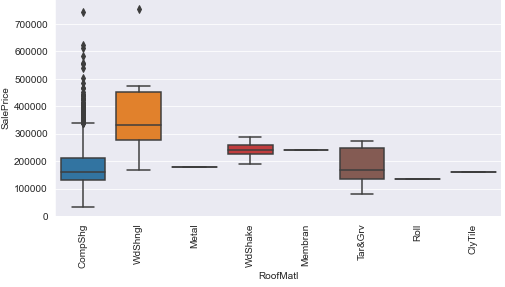
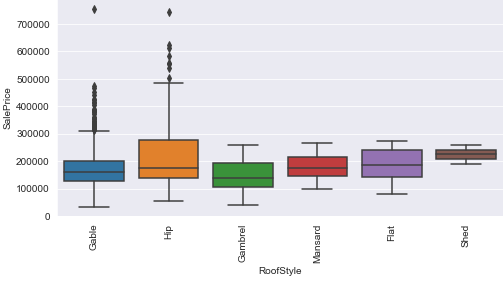


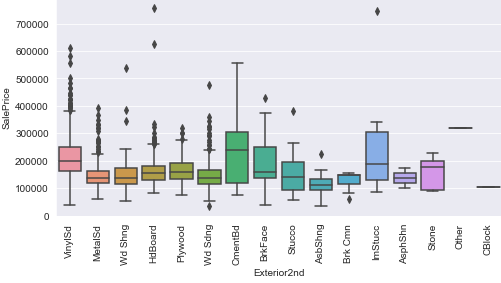
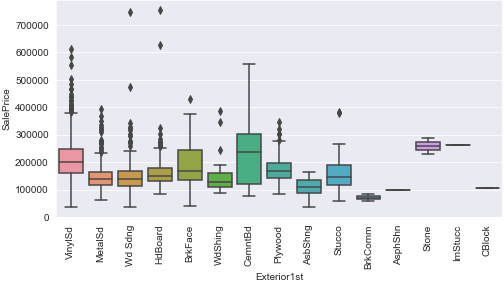


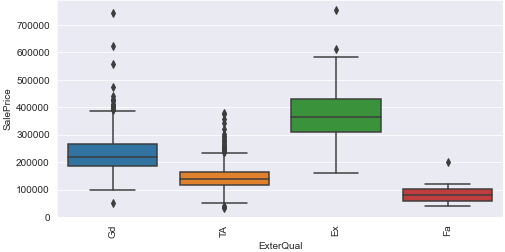
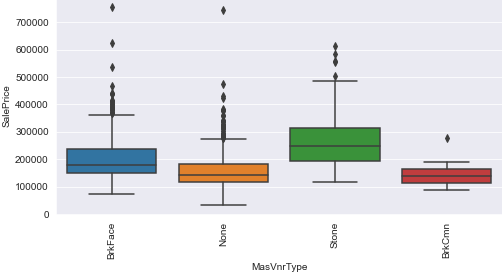


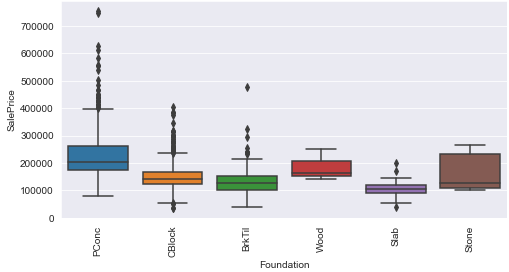
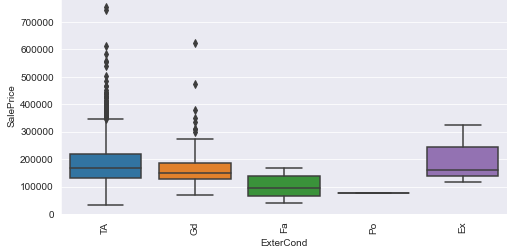


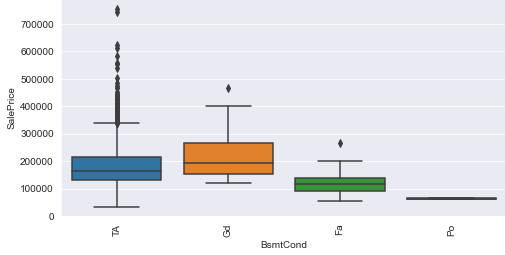
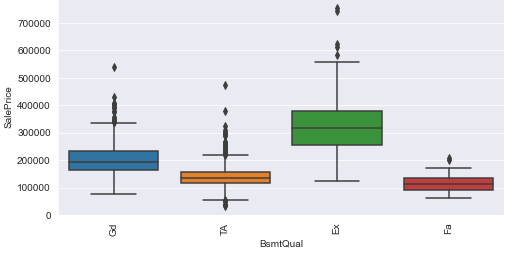


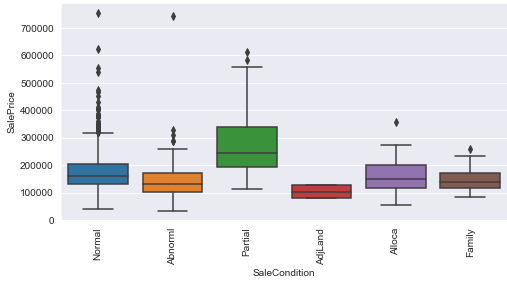
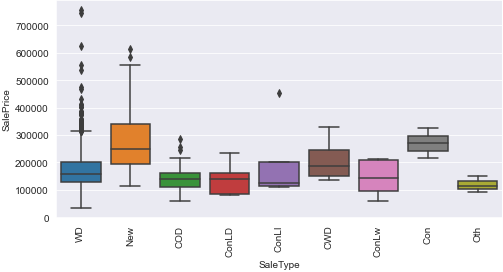
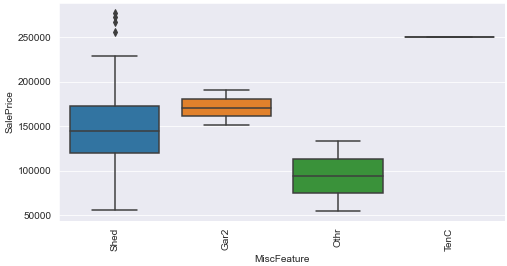
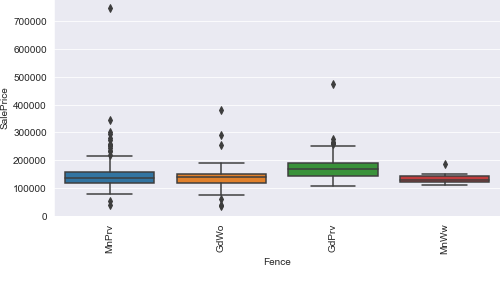
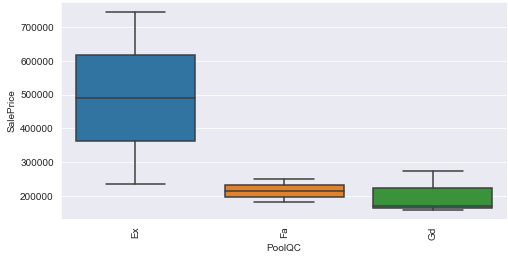
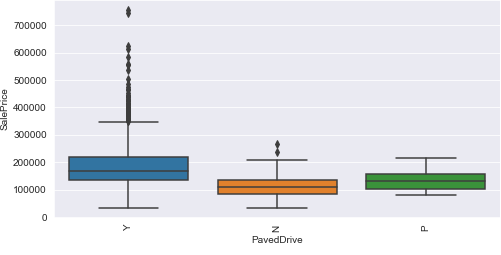
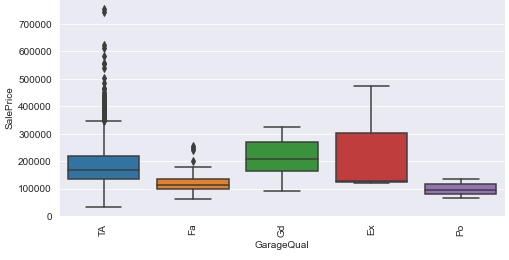
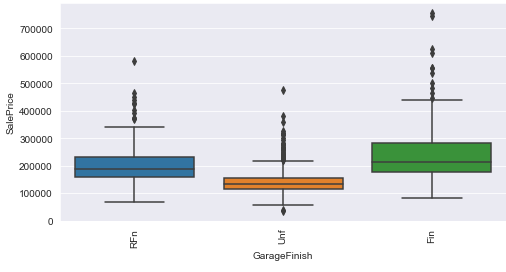
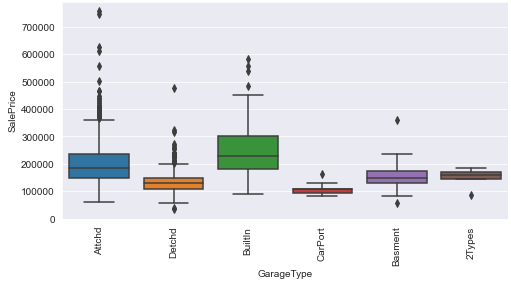
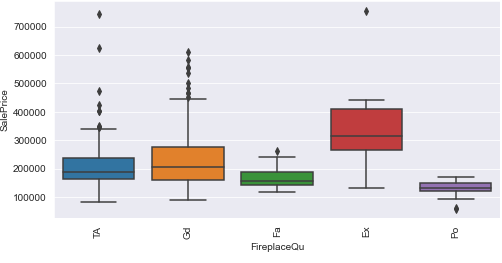
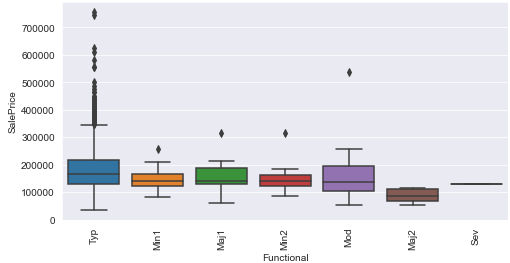
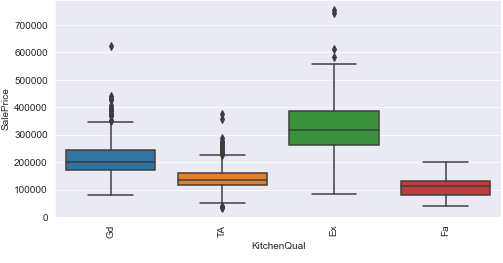
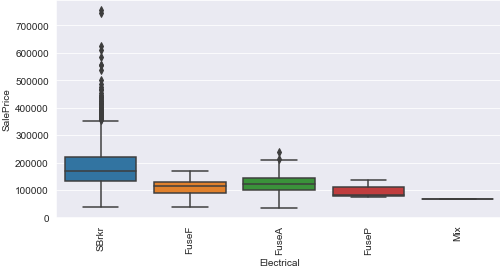
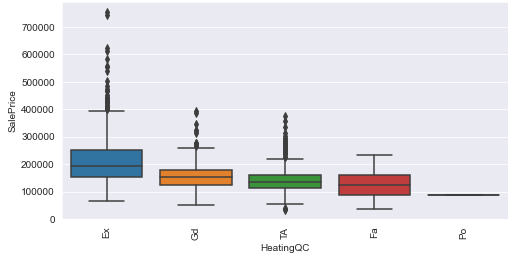
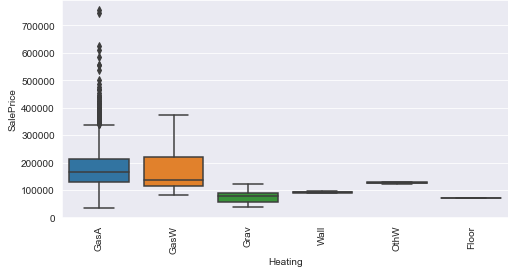
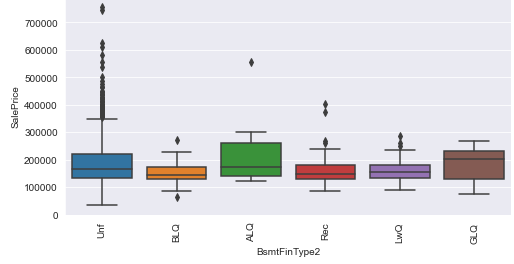
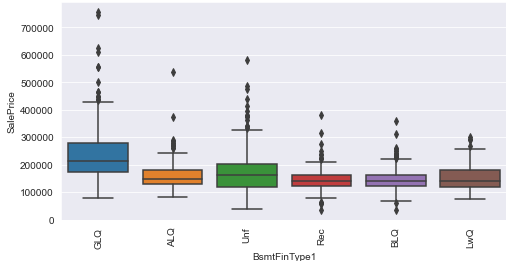
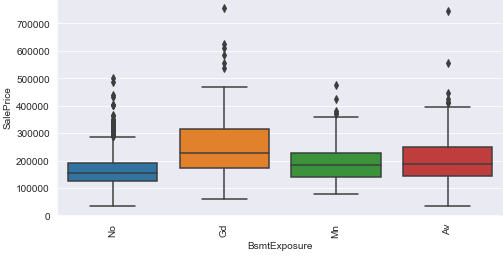






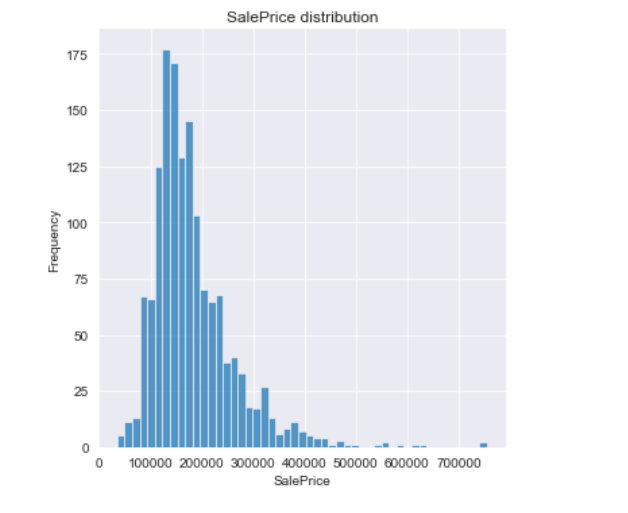






# 

# 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.

# Data Preprocessing :

# 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 |

# 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)*

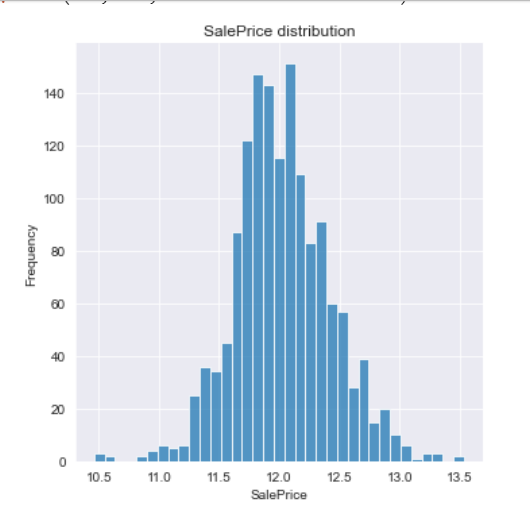
*dataset\_name['PoolQC'] = dataset\_name['PoolQC'] .replace('Po',1, regex=True)*

# Normalise Target value

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

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

Saleprice aftrer distribution

******

# 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.



1. **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)*

1. **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 1.12806 with top 25% in leaderboard.

# 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.