FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:

Prateek Khurana

ACKNOWLEDGMENT

I acknowledge that this Project is done under supervision and guidance of our mentor Mohd. Kashif, data is been provided by him and details of what to do and what not to do is also guided by him.

I am helpful to him without this supervision I will not be able to complete it on time.

INTRODUCTION

- Business Problem Framing

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company..

- Conceptual Background of the Domain Problem

Many of the variable are necessary to get the pricing of a house. In this project I show a relationship of such different variables on final prices. Some of such factors are :- year build up, area, utilities, build quality etc.

- Review of Literature

In this project it is necessary to reflect different factors defining pricing, in the data it has been been seen that there are 76 variables that defines prices so it was necessary t first see ehich of the variables are important and what can be ignored like- id, some variable have direct relationship with prices where some have less impact on prices so I tried in this to project such things and reflect their relationship with prices.

I form different categories which help to understand different variables group like- year group which contain year built, garage year built, year of sale etc and with use of plotmib determine their relationship with final sale price

From this different groups it is easy to determine the final prices in testing model phase for that I used XGBOOST classifier and regressor and randomsearch CV from sklearn.

- Motivation for the Problem Undertaken

It is clear to me in very beginning that I have to frame a model which is capable to project prices based on data provided for this project based on train dataset .

Analytical Problem Framing

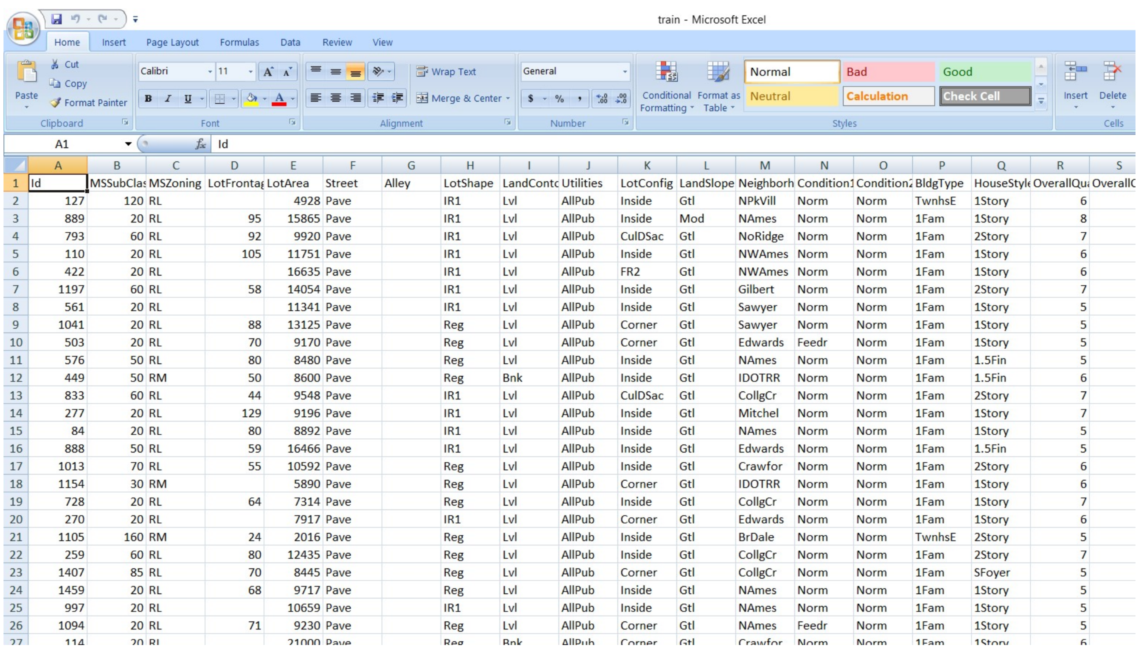
- ## Mathematical/ Analytical Modeling of the Problem

From the data it was seen that many of the variables have missing values, to fill that I have three methods mean, median and mode. I prefer mode in this as that according to me will give best result

- ## Data Sources and their formats

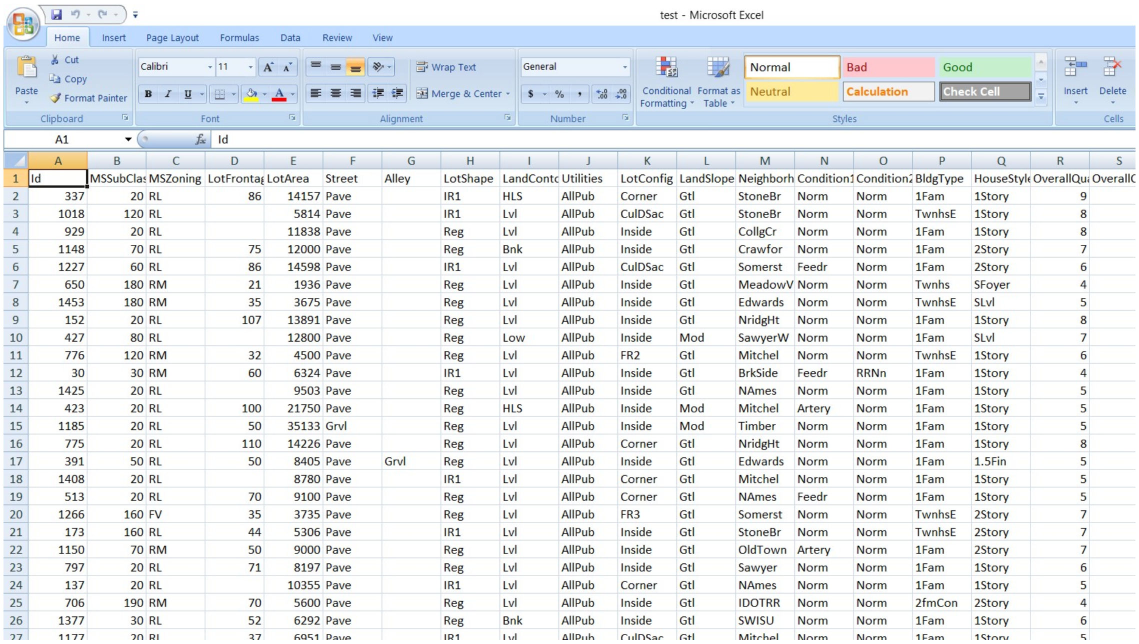
Data is provided by my mentor in two format 1 is train set and other is test set. What I had to do is train the model with train set and then predict result in test set.

Train data:-



	Id	MSSubClas	MSZoning	LotFrontag	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborh	Condition1	Condition2	BldgType	HouseStyl	OverallQu	OverallC
2	127	120	RL		4928	Pave		IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	Norm	TwnhsE	1Story	6	
3	889	20	RL	95	15865	Pave		IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	Norm	1Fam	1Story	8	
4	793	60	RL	92	9920	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	Norm	1Fam	2Story	7	
5	110	20	RL	105	11751	Pave		IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	
6	422	20	RL		16635	Pave		IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	
7	1197	60	RL	58	14054	Pave		IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	
8	561	20	RL		11341	Pave		IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	
9	1041	20	RL	88	13125	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	
10	503	20	RL	70	9170	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	
11	576	50	RL	80	8480	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5Fin	5	
12	449	50	RM	50	8600	Pave		Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	6	
13	833	60	RL	44	9548	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	
14	277	20	RL	129	9196	Pave		IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	7	
15	84	20	RL	80	8892	Pave		IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	
16	888	50	RL	59	16466	Pave		IR1	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1.5Fin	5	
17	1013	70	RL	55	10592	Pave		Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	6	
18	1154	30	RM		5890	Pave		Reg	Lvl	AllPub	Corner	Gtl	IDOTRR	Norm	Norm	1Fam	1Story	6	
19	728	20	RL	64	7314	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	
20	270	20	RL		7917	Pave		IR1	Lvl	AllPub	Corner	Gtl	Edwards	Norm	Norm	1Fam	1Story	6	
21	1105	160	RM	24	2016	Pave		Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	TwnhsE	2Story	5	
22	259	60	RL	80	12435	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	
23	1407	85	RL	70	8445	Pave		Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	SFoyer	5	
24	1459	20	RL	68	9717	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	
25	997	20	RL		10659	Pave		IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	
26	1094	20	RL	71	9230	Pave		Reg	Lvl	AllPub	Corner	Gtl	NAmes	Feedr	Norm	1Fam	1Story	5	
27	114	20	RL		21000	Pave		Reg	Bnk	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	1Story	6	

Test data:-



Id	MSSubClas	MSZoning	LotFrontag	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborh	Condition1	Condition2	BldgType	HouseStyle	OverallQu	OverallC
337	20	RL	86	14157	Pave		IR1	HLS	AllPub	Corner	Gtl	StoneBr	Norm	Norm	1Fam	1Story	9	
1018	120	RL		5814	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	8	
929	20	RL		11838	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	
1148	70	RL	75	12000	Pave		Reg	Bnk	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	
1227	60	RL	86	14598	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	Somerst	Feedr	Norm	1Fam	2Story	6	
650	180	RM	21	1936	Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	Twnhs	SFoyer	4	
1453	180	RM	35	3675	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	TwnhsE	SLvl	5	
152	20	RL	107	13891	Pave		Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	
427	80	RL		12800	Pave		Reg	Low	AllPub	Inside	Mod	SawyerW	Norm	Norm	1Fam	SLvl	7	
776	120	RM	32	4500	Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	TwnhsE	1Story	6	
30	30	RM	60	6324	Pave		IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr	RRNn	1Fam	1Story	4	
1425	20	RL		9503	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	
423	20	RL	100	21750	Pave		Reg	HLS	AllPub	Inside	Mod	Mitchel	Artery	Norm	1Fam	1Story	5	
1185	20	RL	50	35133	Grvl		Reg	Lvl	AllPub	Inside	Mod	Timber	Norm	Norm	1Fam	1Story	5	
775	20	RL	110	14226	Pave		Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	
391	50	RL	50	8405	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1.5Fin	5	
1408	20	RL		8780	Pave		IR1	Lvl	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	5	
513	20	RL	70	9100	Pave		Reg	Lvl	AllPub	Corner	Gtl	NAmes	Feedr	Norm	1Fam	1Story	5	
1266	160	FV	35	3735	Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	
173	160	RL	44	5306	Pave		IR1	Lvl	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	2Story	7	
1150	70	RM	50	9000	Pave		Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	2Story	7	
797	20	RL	71	8197	Pave		Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	6	
137	20	RL		10355	Pave		IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	
706	190	RM	70	5600	Pave		Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	2fmCon	2Story	4	
1377	30	RL	52	6292	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	
1177	20	RL	37	6951	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	1Story	5	

Here we check the percentage of nan values present in each feature :-

LotFrontage 0.1832 %
Alley 0.9341 %
MasVnrType 0.006 %
MasVnrArea 0.006 %
BsmtQual 0.0257 %
BsmtCond 0.0257 %
BsmtExposure 0.0265 %
BsmtFinType1 0.0257 %
BsmtFinType2 0.0265 %
FireplaceQu 0.4717 %
GarageType 0.0548 %
GarageYrBlt 0.0548 %
GarageFinish 0.0548 %
GarageQual 0.0548 %
GarageCond 0.0548 %
PoolQC 0.994 %
Fence 0.7971 %

MiscFeature 0.9623 %

- Data Preprocessing Done

Filling the missing value we need to see the test and train data simultaneously.
We will be replacing the null values with mode for categorical values, discrete numerical values and year variables
We will be replacing the null values with mean for continuous numerical values.
We will delete columns with more than 50% null values as the available information add no value for our model.

For that first check null value with use of heat map

```
sns.heatmap(df.isnull(),yticklabels = False,cbar = False)
```

after knowing now filling values

```
for feature in categorical_features:
    df[feature] = df[feature].fillna(df[feature].mode()[0])
#train
    df_test[feature] =
df_test[feature].fillna(df_test[feature].mode()[0])  #test

for feature in discrete_features:
    df[feature] = df[feature].fillna(df[feature].mode()[0])
#train
    df_test[feature] =
df_test[feature].fillna(df_test[feature].mode()[0])  #test

for feature in year_features:
```

```
    df[feature] = df[feature].fillna(df[feature].mode()[0])
#train
    df_test[feature] =
df_test[feature].fillna(df_test[feature].mode()[0])  #test


for feature in continous_numerical_features:
    df[feature] = df[feature].fillna(df[feature].mean())
#train
    df_test[feature] =
df_test[feature].fillna(df_test[feature].mean())  #test
```

Now after filling missing values I drop columns that does not impact much on final results:

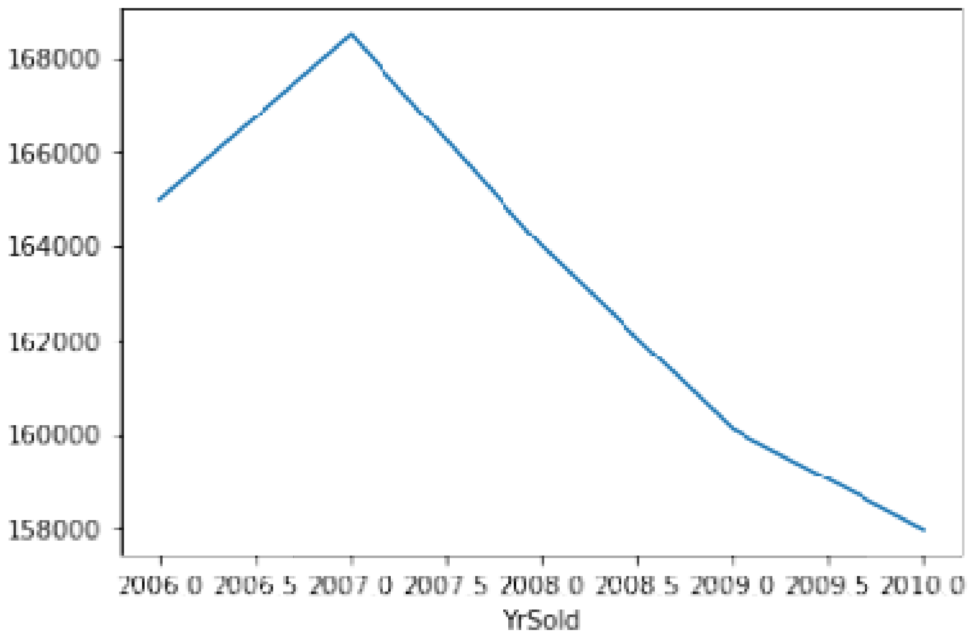
```
for feature in
more_than_50_percent_misssing_value_features:
    df.drop([feature],axis = 1, inplace = True)
    df_test.drop([feature],axis = 1, inplace = True)


df.drop(['Id'],axis = 1, inplace = True)
df_test.drop(['Id'],axis = 1, inplace = True
```

- Data Inputs- Logic- Output Relationships

In the given data there are many variables that can impact the output such as year, quality, layout, area, etc. and each factor is capable to impact the sale price.

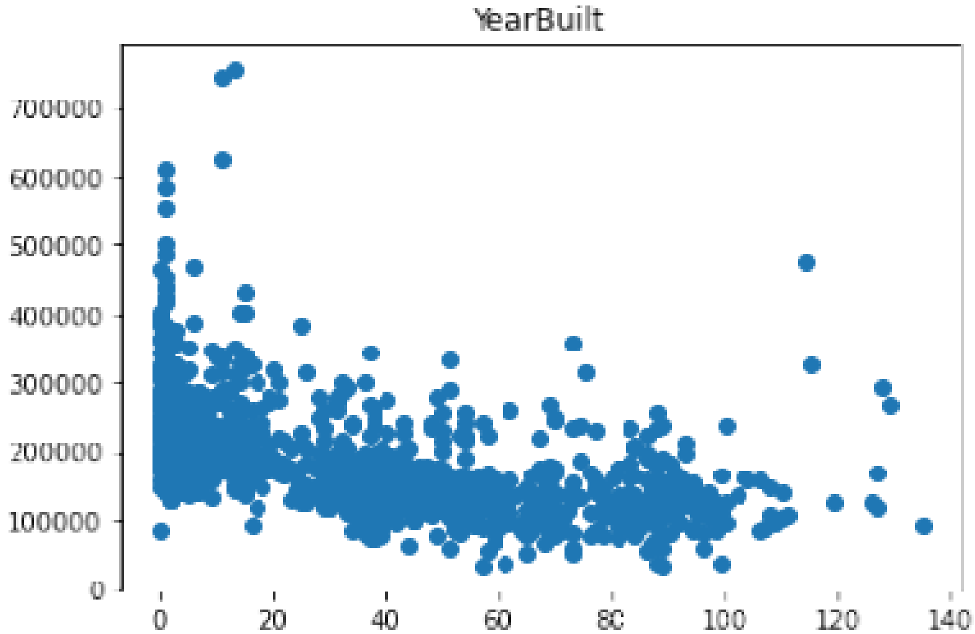
Year group:

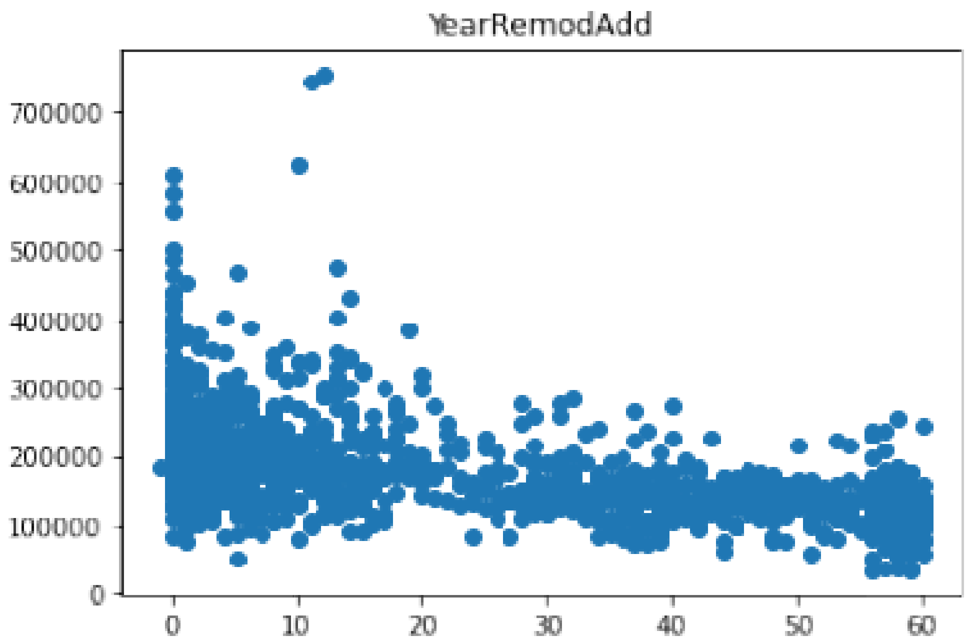
df.groupby('YrSold')['SalePrice'].median().plot()

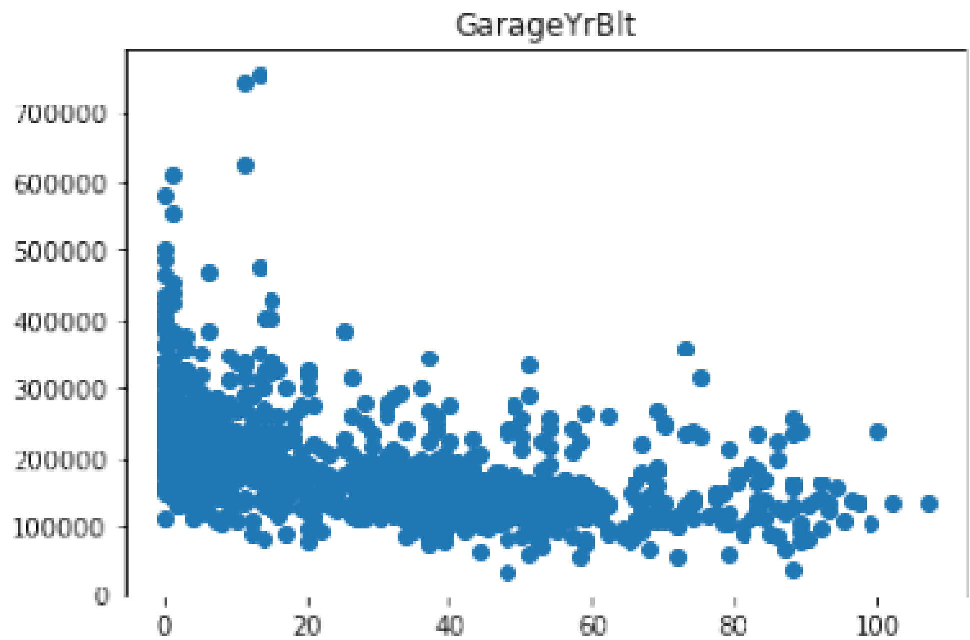


In year froup

```
for feature in year_features:
    if feature != "YrSold":
        data = df.copy()
        data[feature] = data['YrSold'] - data[feature]
        plt.scatter(data[feature],data['SalePrice'])
        plt.title(feature)
        plt.show()
```


YearBuilt
700000
600000
500000
400000
300000
200000
100000
0
0
20
40
60
80
100
120
140


YearRemodAdd
700000
600000
500000
400000
300000
200000
100000
0
0
10
20
30
40
50
60



Model/s Development and Evaluation

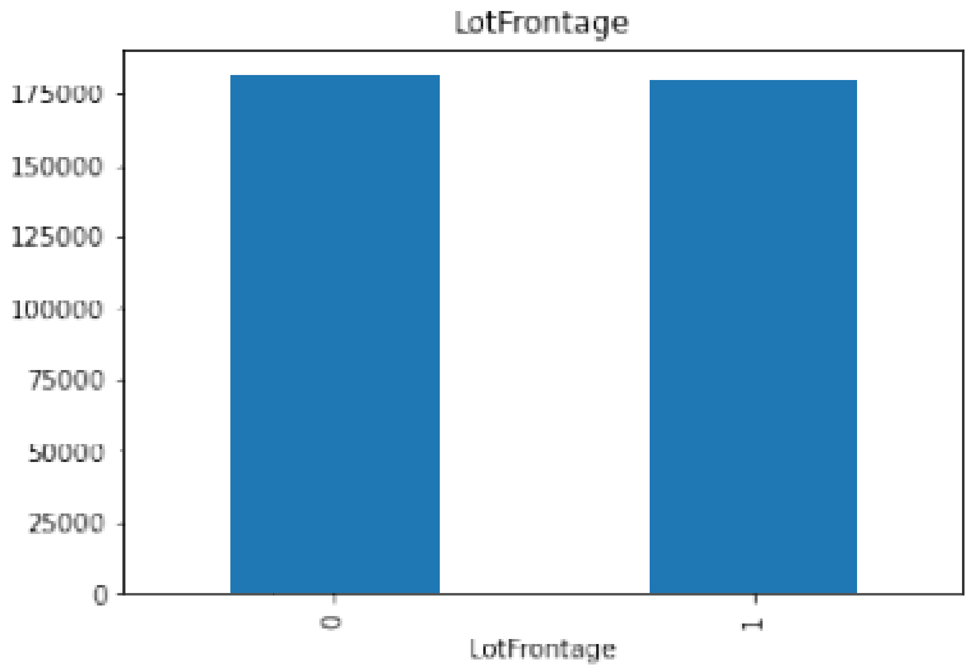
Testing of Identified Approaches (Algorithms)

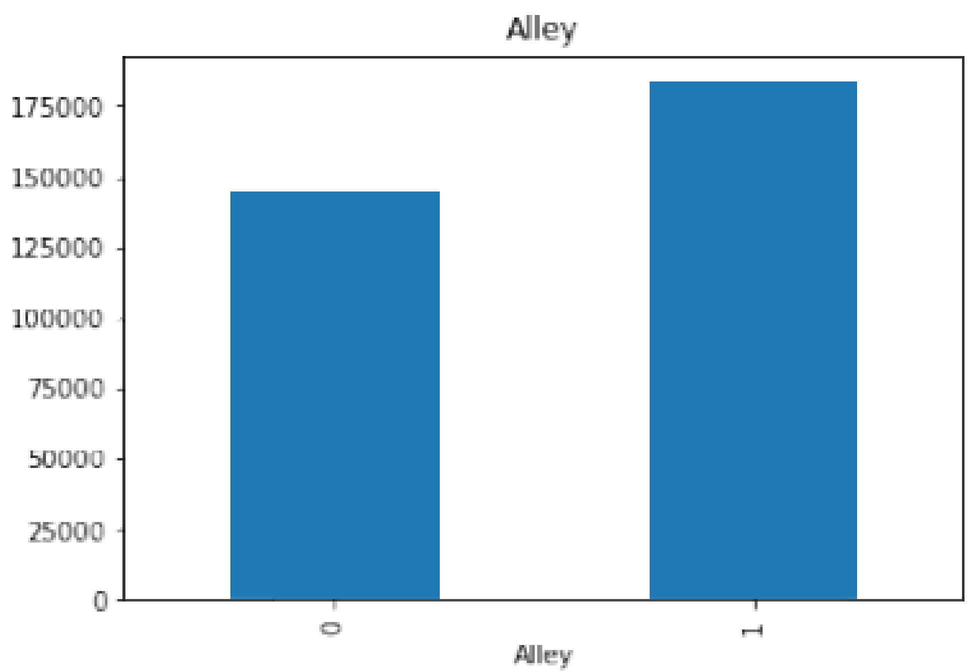
XGboost is the most widely used algorithm in machine learning, whether the problem is a classification or a regression problem. It is known for its good performance as compared to all other machine learning algorithms.

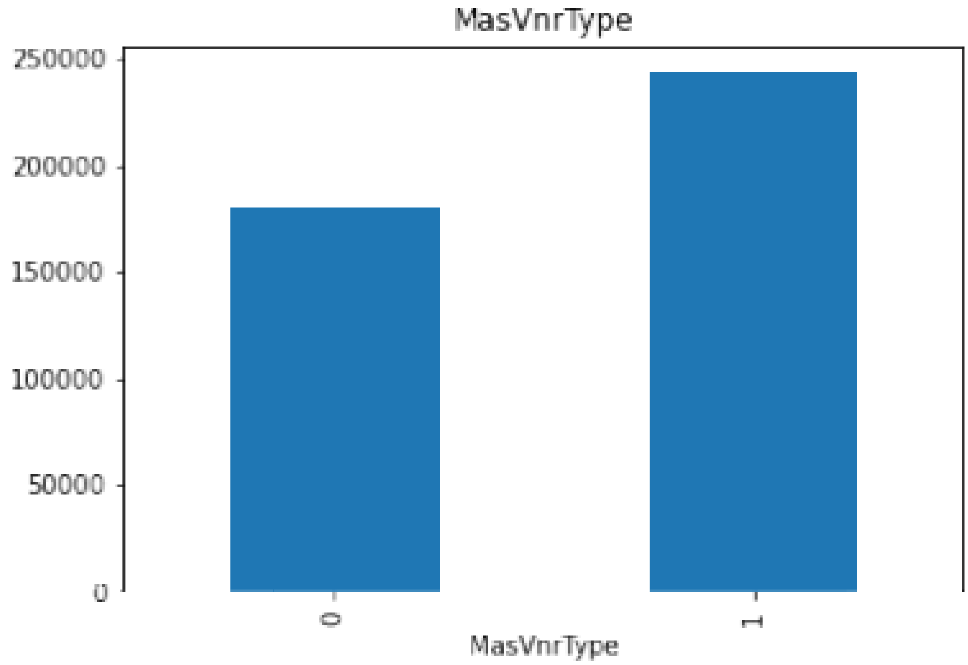
Even when it comes to machine learning competitions and hackathon, XGBoost is one of the excellent algorithms that is picked initially for structured data. It has proved its determination in terms of speed and performance.

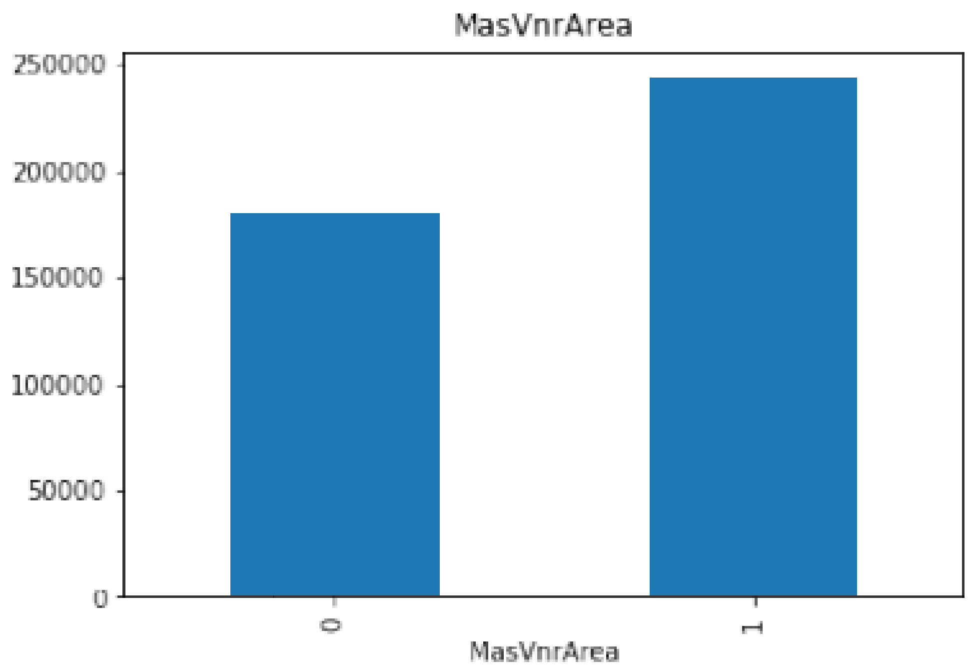
.

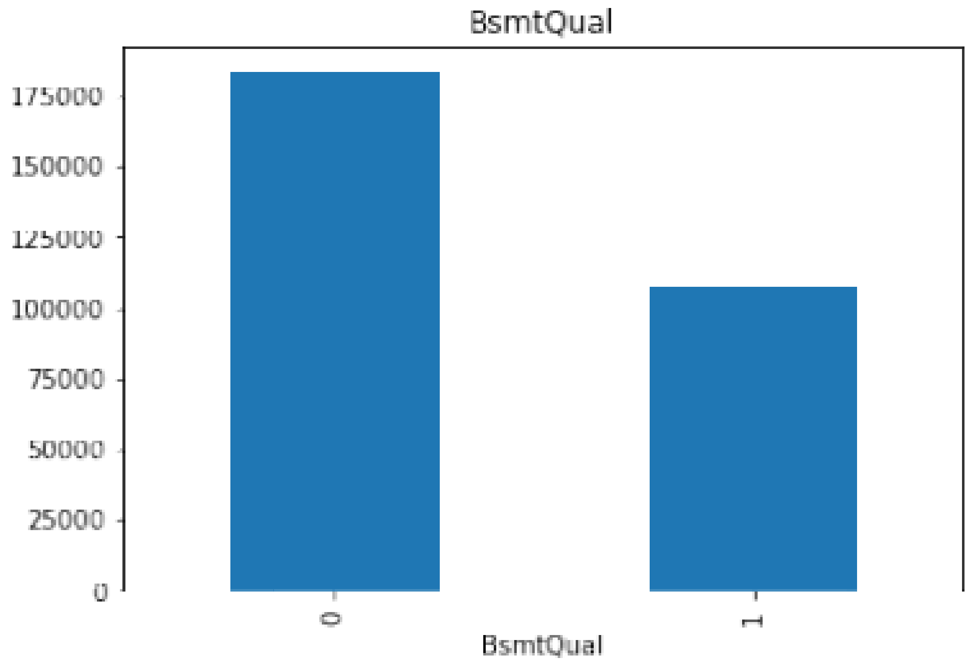
- Visualizations
  For visualizations I used metplot feature and in that I used bar plot for missing values and features having less than 50% impact on result

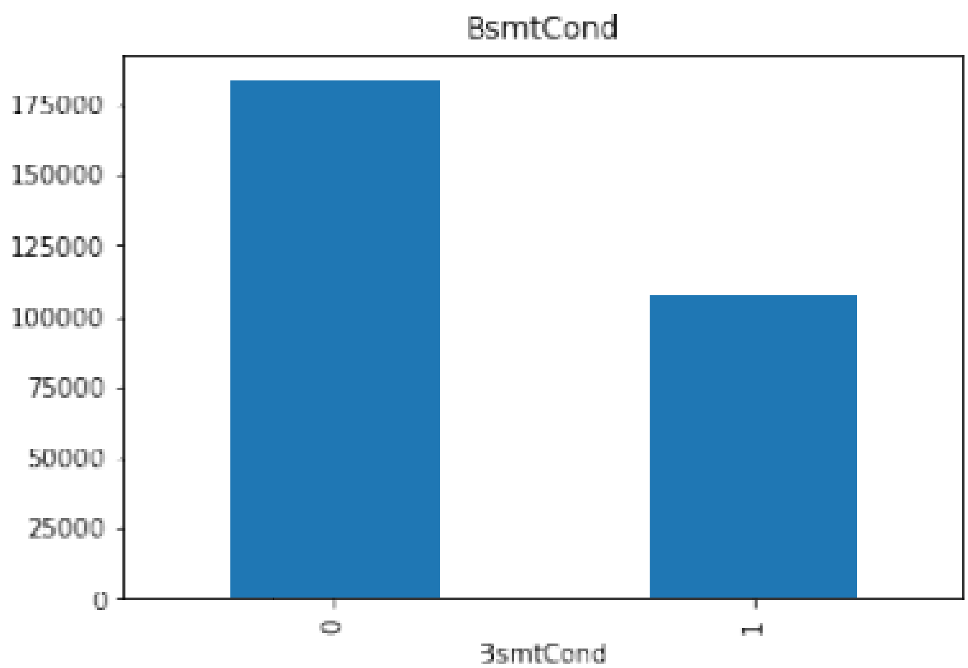


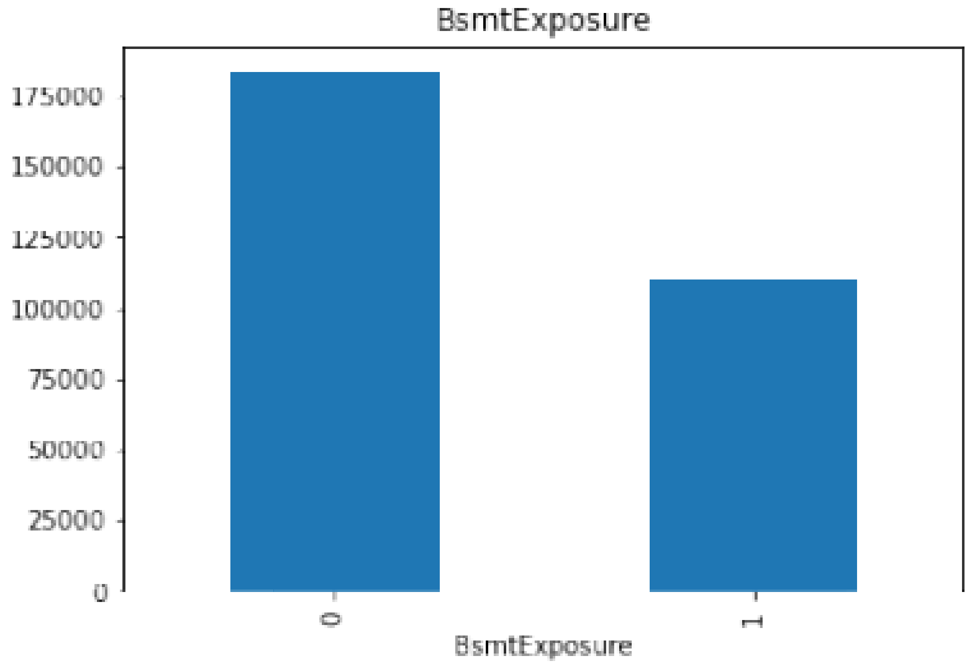


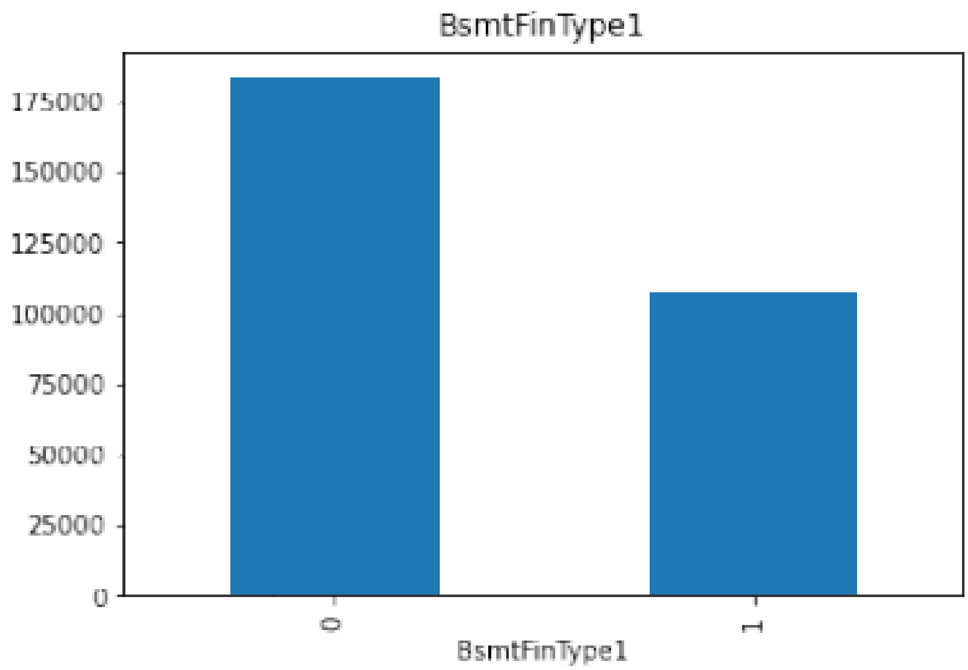

MasVnrType
250000
200000
150000
100000
50000
0
0
1
MasVnrType

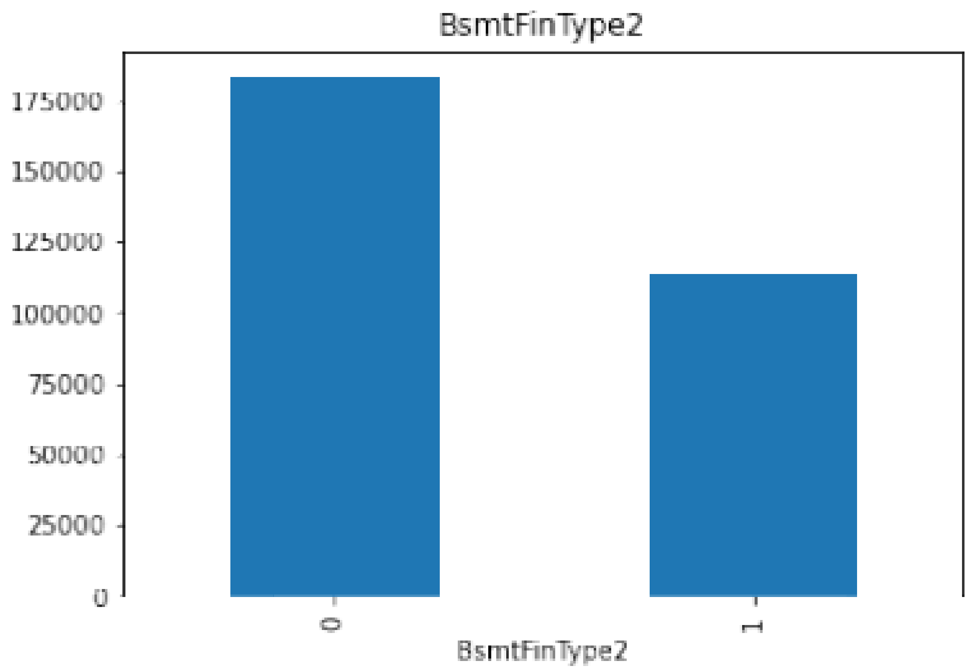

MasVnrArea
250000
200000
150000
100000
50000
0
0
1
MasVnrArea


BsmtQual
175000
150000
125000
100000
75000
50000
25000
0
0
1
BsmtQual

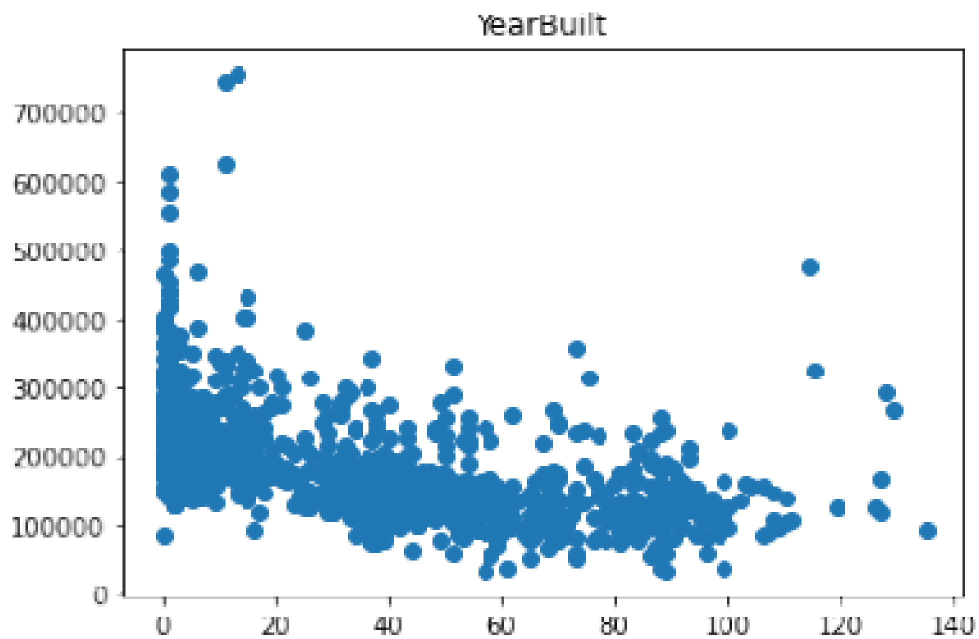

BsmtCond
175000
150000
125000
100000
75000
50000
25000
0
0
1
BsmtCond


BsmtExposure
175000
150000
125000
100000
75000
50000
25000
0
0
1
BsmtExposure

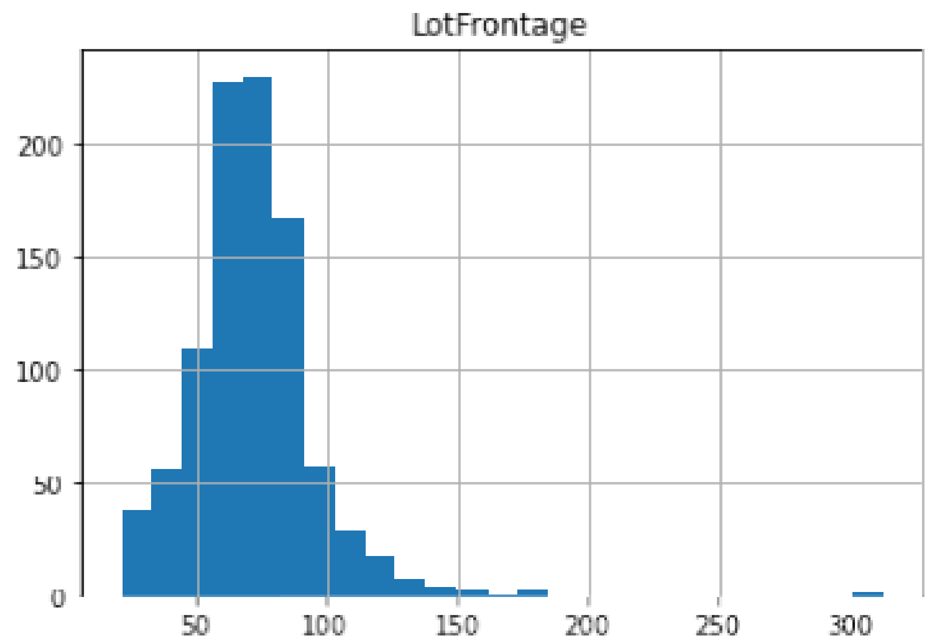

BsmtFinType1
175000
150000
125000
100000
75000
50000
25000
0
0
1
BsmtFinType1



Then used scatter plot for year group



Then used histogram plot for continuous feature



CONCLUSION

- Key Findings and Conclusions of the Study

From whole analysis of given data through different means like metplot and Xgboost, it made easy to understand the impact each variable have on Sale price like garage on sale price, area on sale price. With these tools help I am able to built a model that can predict final sale price for test data for me

- Limitations of this work and Scope for Future Work

In this project data were having many missing values hence the result obtained can't be 100% true, it have its limitations as I used mode method to fill missing values even though result will deflect different values if used mean or median method.