HOUSE PRICE PREDICTION

PARTH PATEL

Who might care?

Real estate investors

Banks

Local home buyers and local seller

Data Overview

- Data set obtained from Kaggle
- Number of rows ~3K
- Column Description
  - About 80 features
  - Data include numerical, categorical and time based data
  - Features based on different sections of house and their attribute

MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig
60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside
20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2
60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside
70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner
60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2
50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	Inside
20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	Inside
60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	Corner
50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	Inside
190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	Corner
20	RL	70.0	11200	Pave	NaN	Reg	Lvl	AllPub	Inside
60	RL	85.0	11924	Pave	NaN	IR1	Lvl	AllPub	Inside
20	RL	NaN	12968	Pave	NaN	IR2	Lvl	AllPub	Inside
20	RL	91.0	10652	Pave	NaN	IR1	Lvl	AllPub	Inside

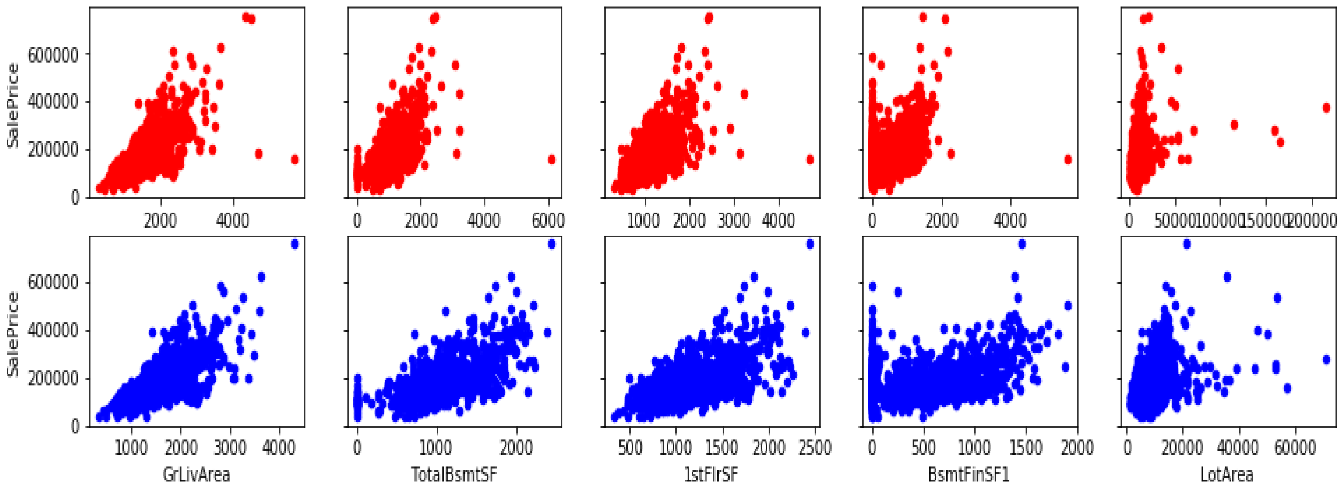
Data Cleaning and Outlier

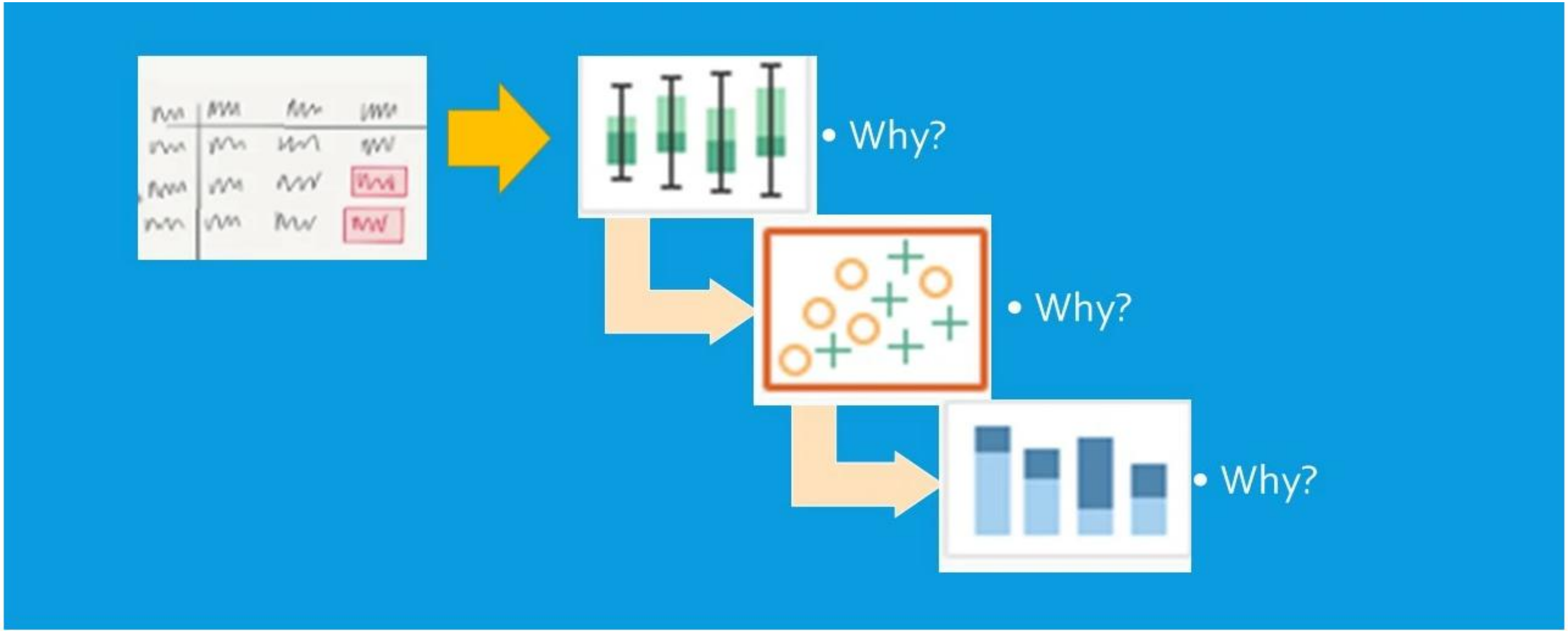
Data Cleaning and Outlier detection

- Features with basement and garage were replaced with 0 or Missing as they represent absence of feature
- Feature like Pool quality, Miscellaneous Feature , alley and fence were dropped
- Lot Frontage was replaced with mean grouped by neighborhood
- Missing Functional were of type Typ as per documentation

	Total	Percent
SalePrice	1459	49.982871
LotFrontage	486	16.649538
GarageYrBlt	159	5.447071
MasVnrArea	23	0.787941
BsmtHalfBath	2	0.068517
BsmtFullBath	2	0.068517

	Total	Percent
PoolQC	2909	99.657417
MiscFeature	2814	96.402878
Alley	2721	93.216855
Fence	2348	80.438506
FireplaceQu	1420	48.646797
GarageCond	159	5.447071

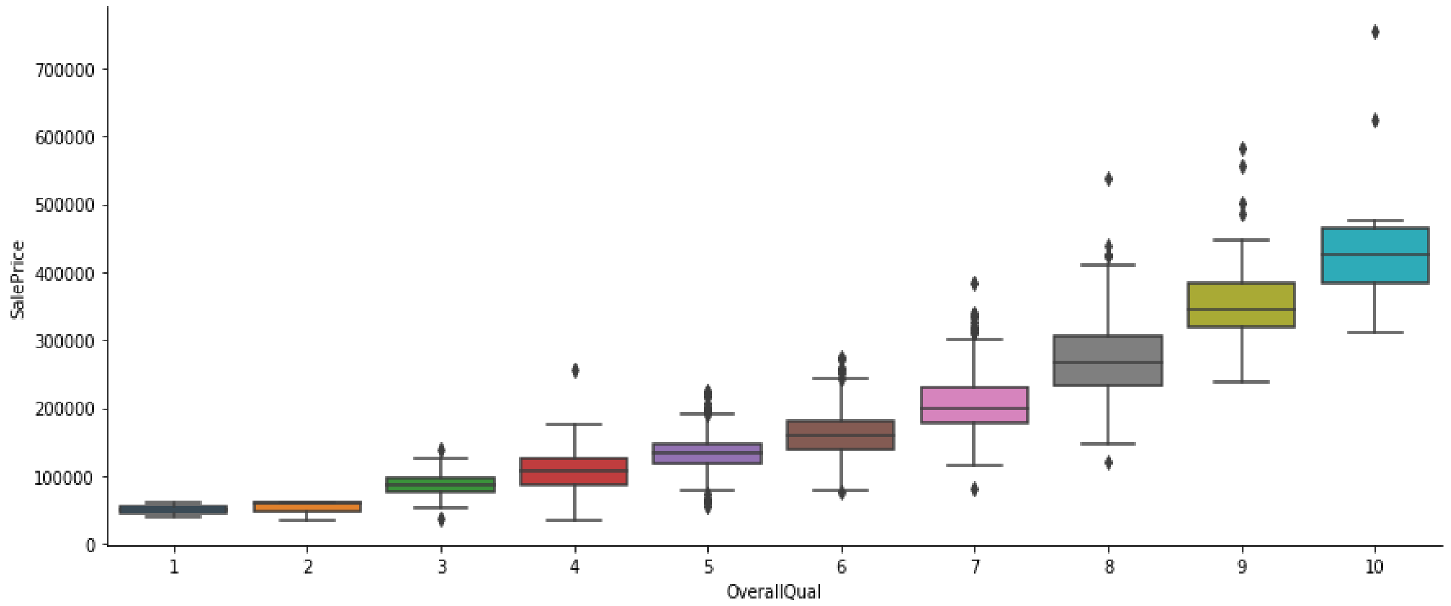


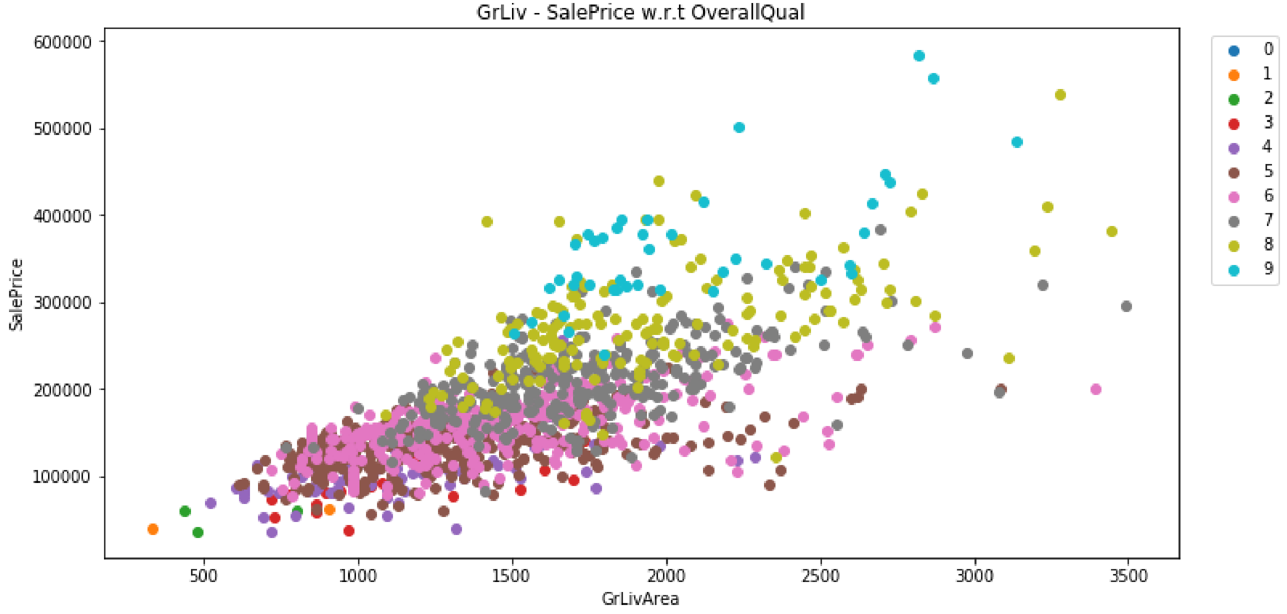


Exploratory Data Analysis

Overall Quality and Living area

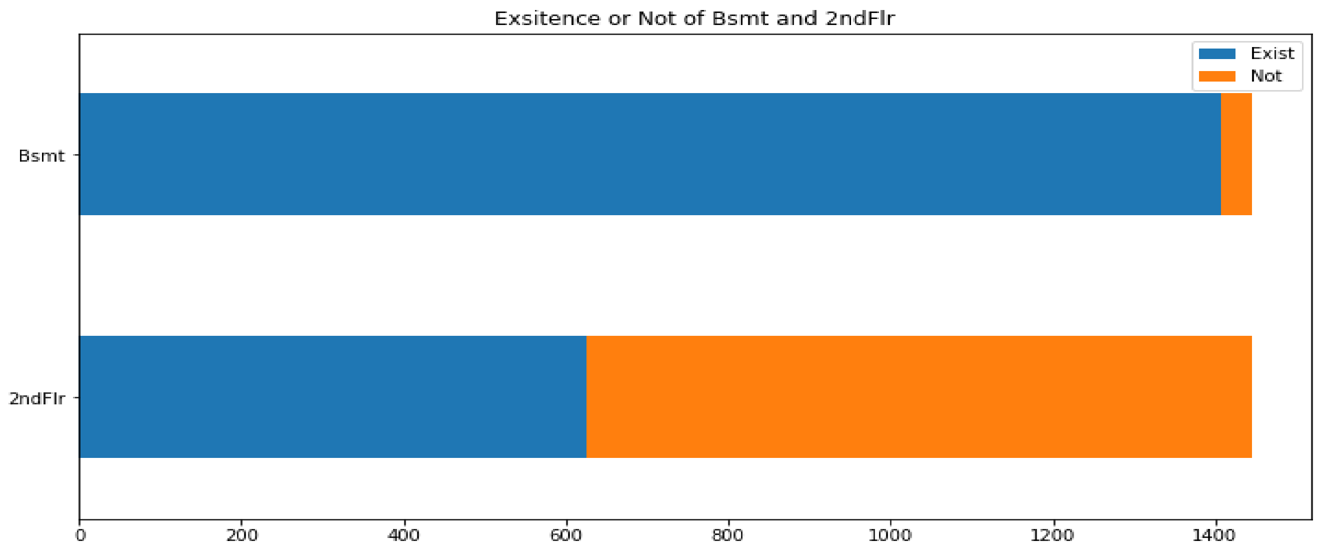
- Overall Quality is the very good variables to explaining Sale Price
- Overall Quality causes different Sale Price where having same "GrLivArea".
- Overall Quality was proportional to SalePrice

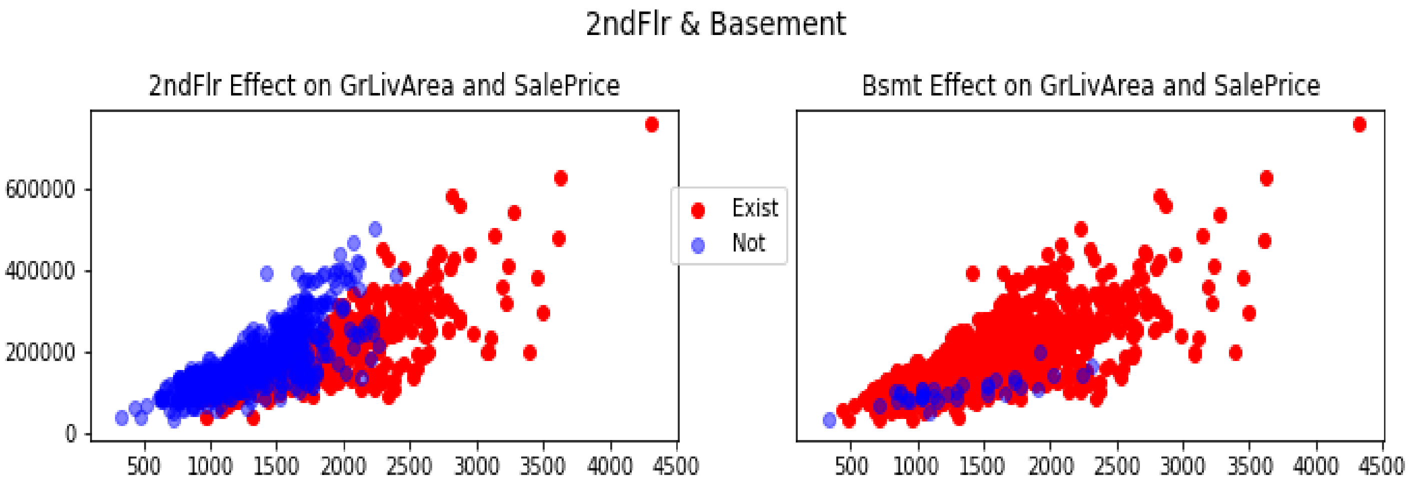




Basement and Upper floor

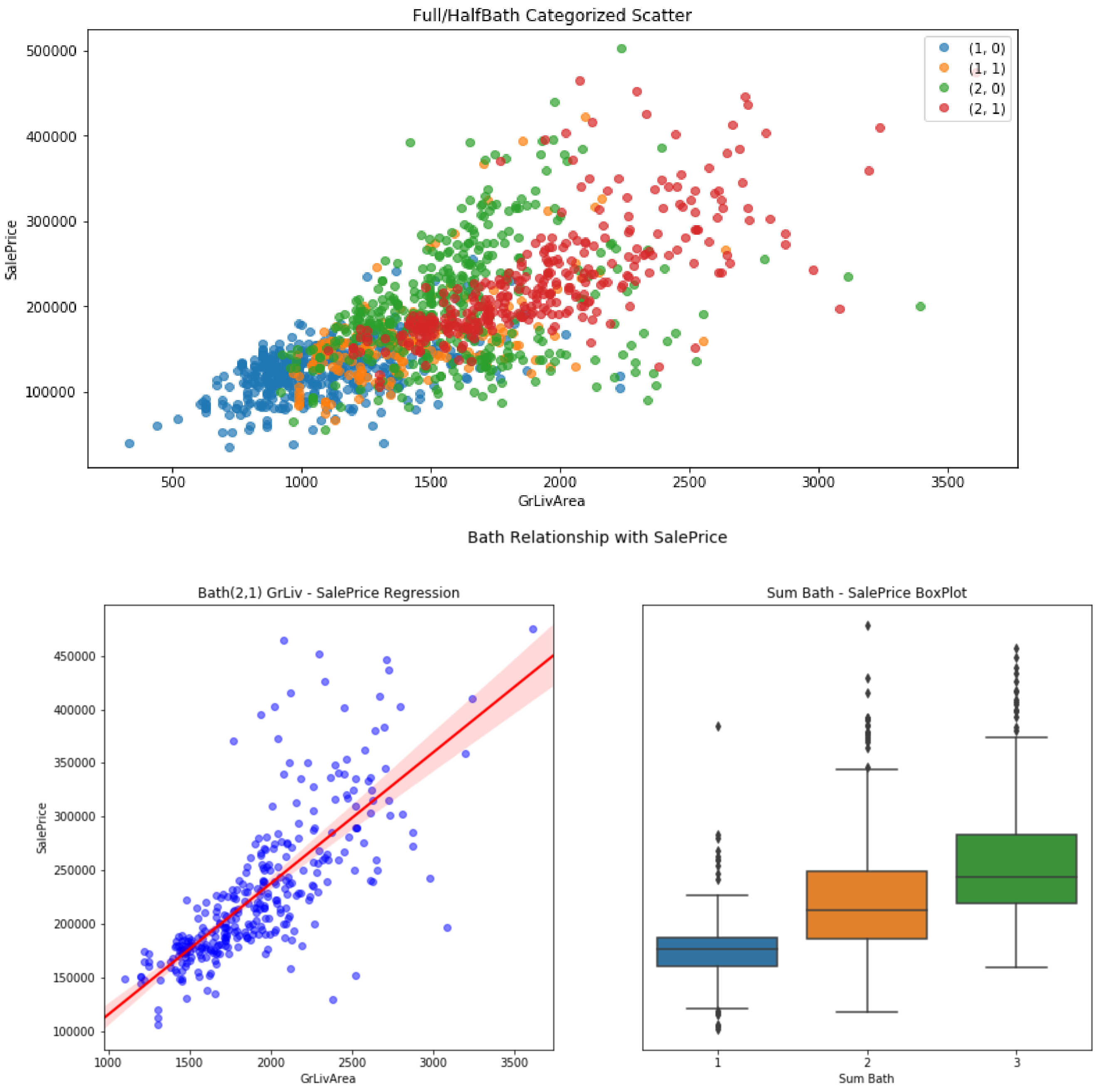
- 2ndFlrSF depressed the power of GrLivArea toward Sale Price
- Basement has nothing related to the price





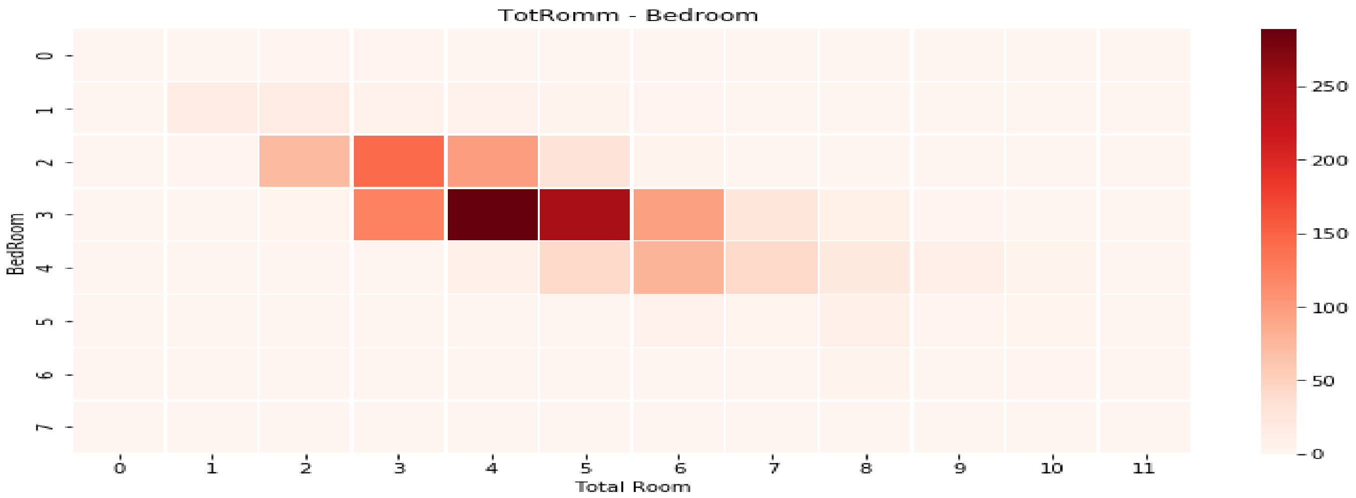
Bathroom

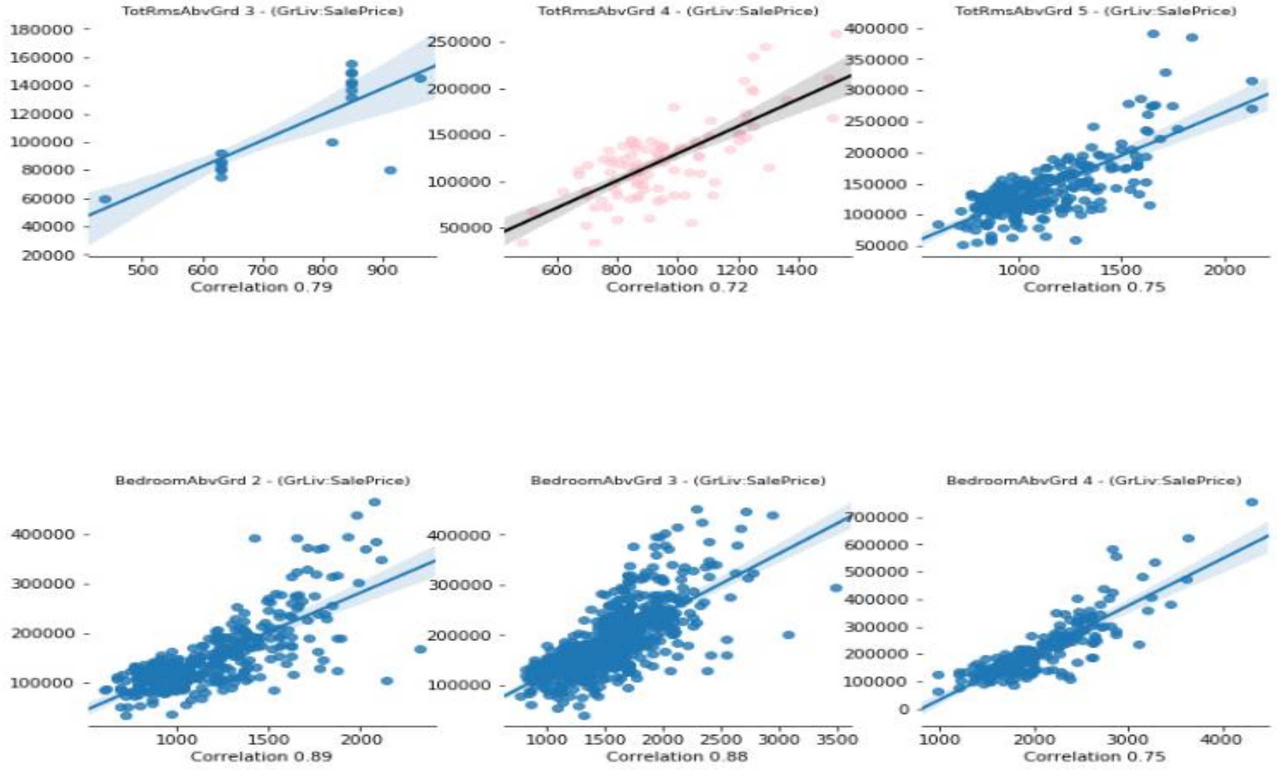
- The Number of Bath usually increased the Sale Price
- combination of (Full 2, Half1) improved the linearity and decreased the Spreadness of Sale Price - GrLivArea.



Technical Room

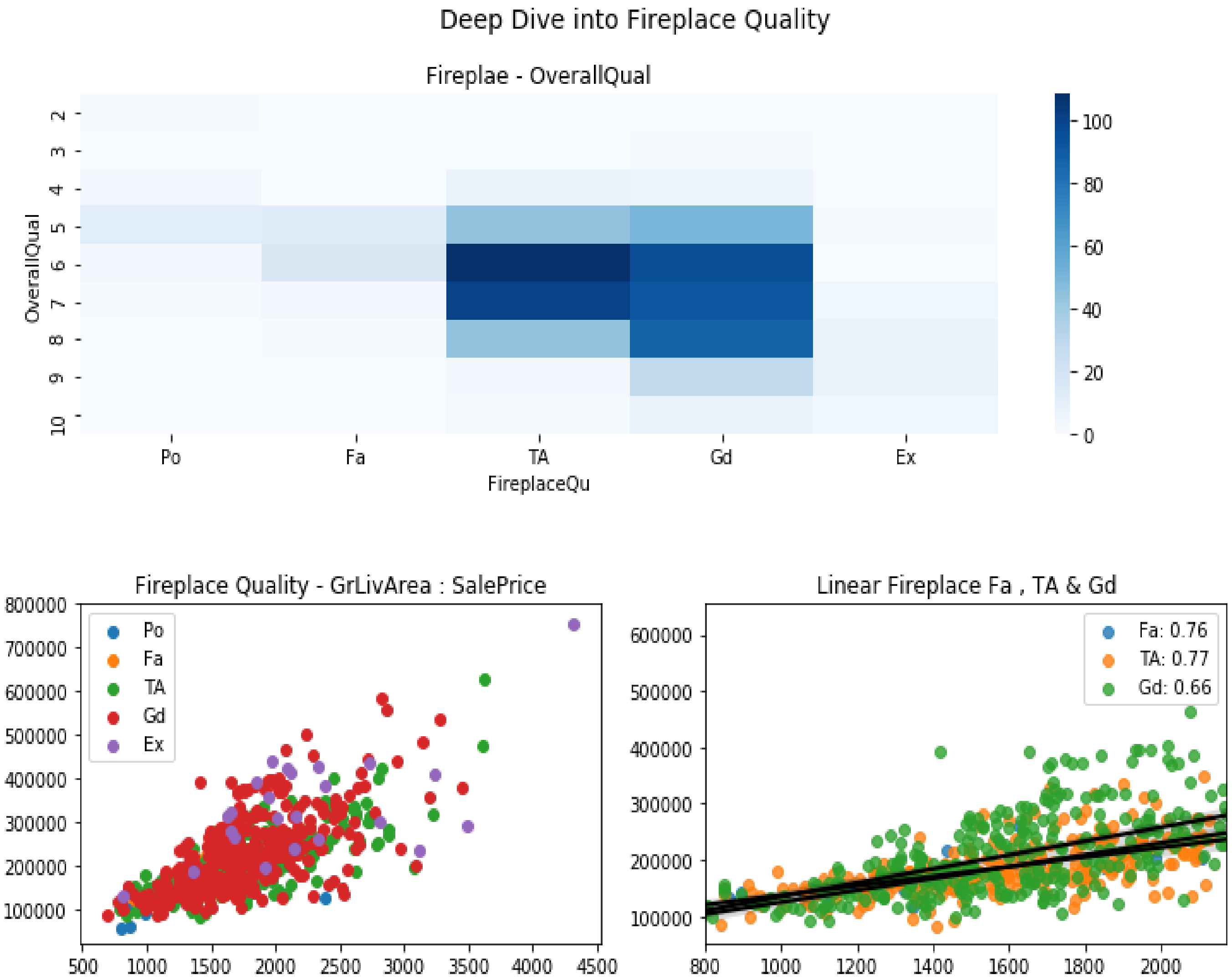
- Total room above ground and no of bedroom are linearly related
- total room above ground also have very good correlation above 0.7
- bedroom has very high correlation with sale price and living area(usually above 0.75)





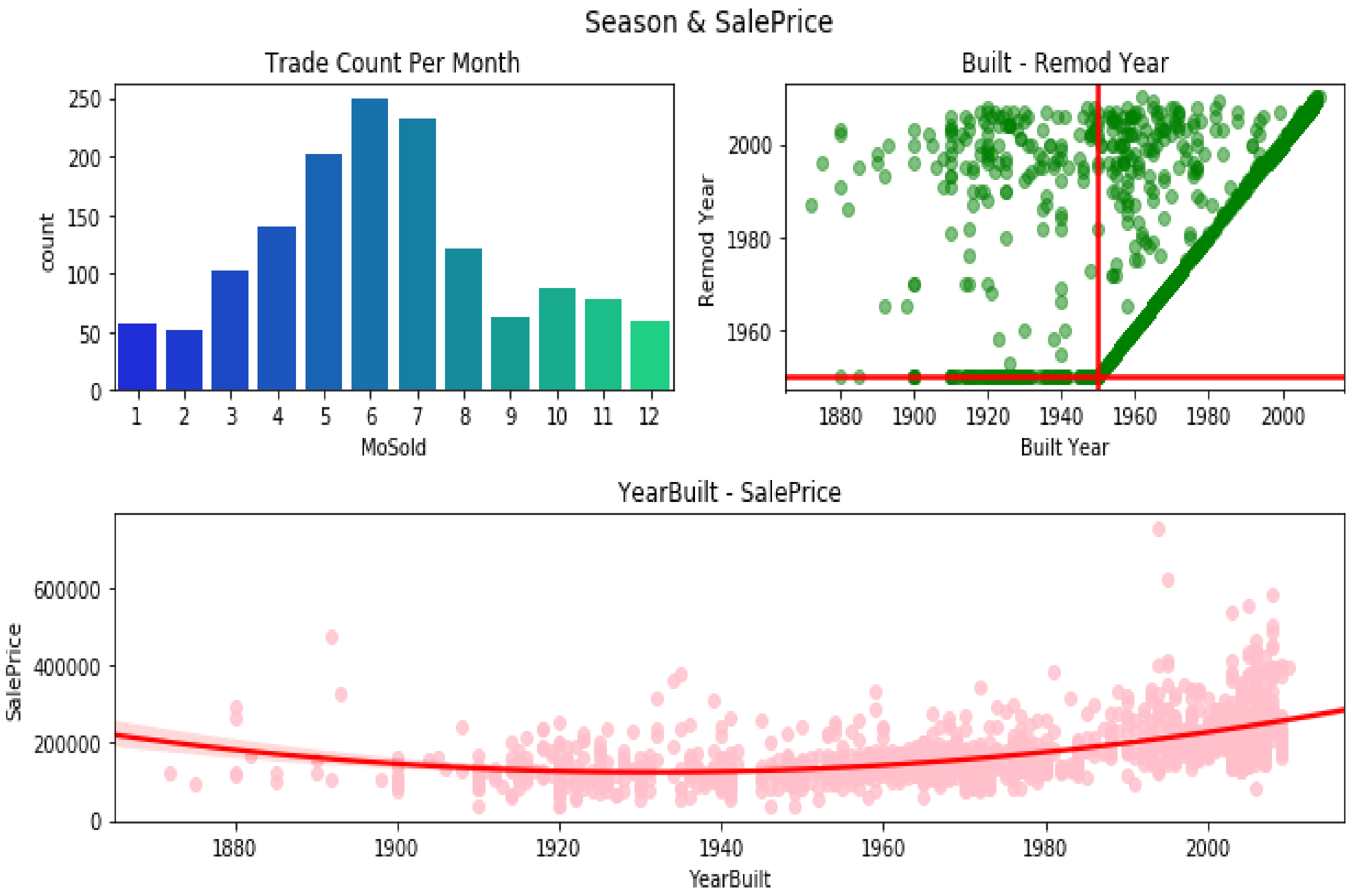
Outside area

- Good Quality House has more outside instrumental places.
- Pool Area, Screen Porch, 3SsnPorch were almost negligible thus not plotted
- Fireplace is linear related to overall quality
- only Fa,TA and Gd has some linear relationship
- Ex and Po does not have good linear relationship, which is clearly visible by looking where Ex are so spread around



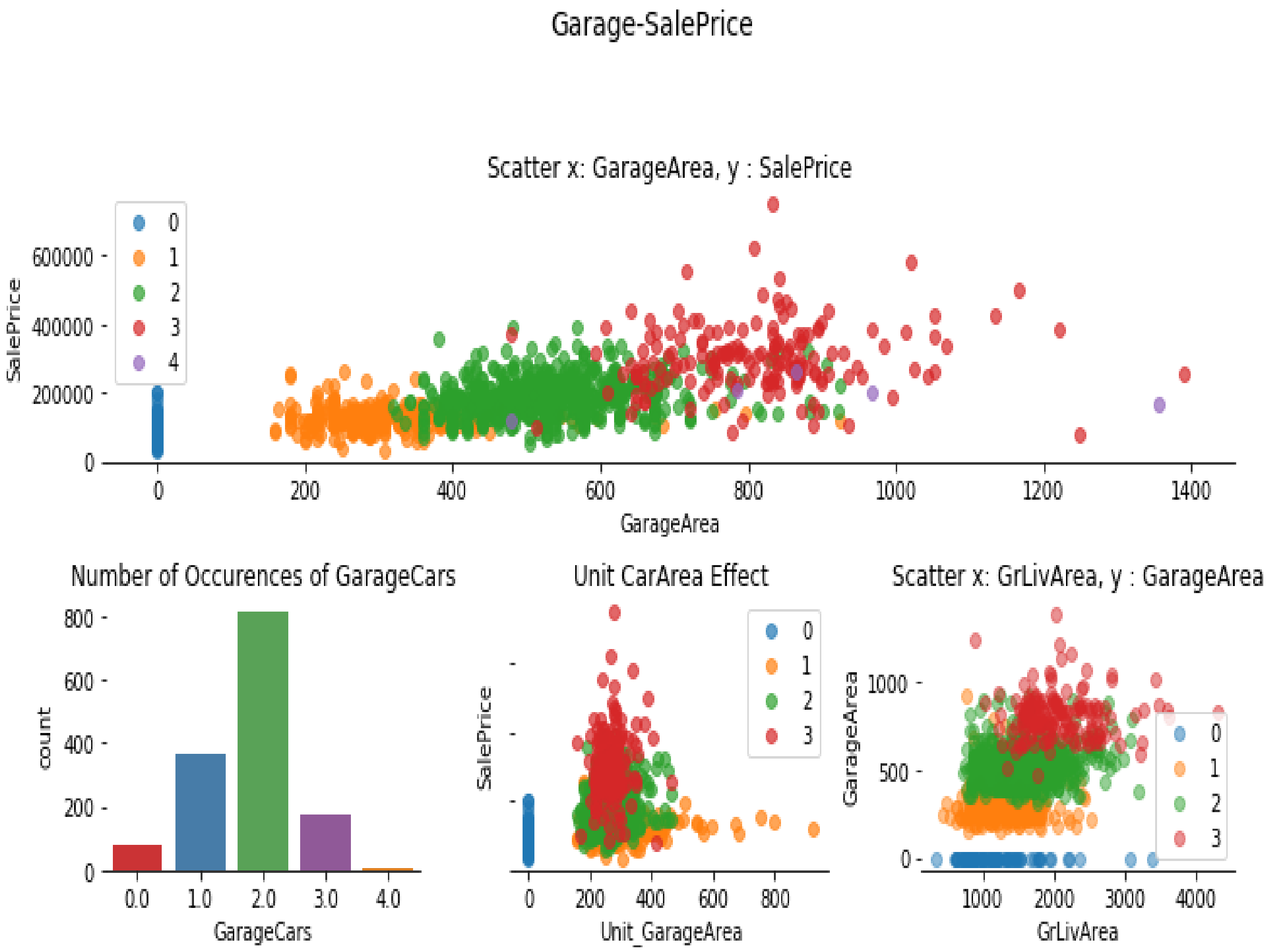
Effect of Season

- The amount of trade was increased by rising temperature(Less trend in winter)
- The part of house, built after 1950, was not remodeled yet
- YearBuilt^2 will be proper if the variables is used to predict



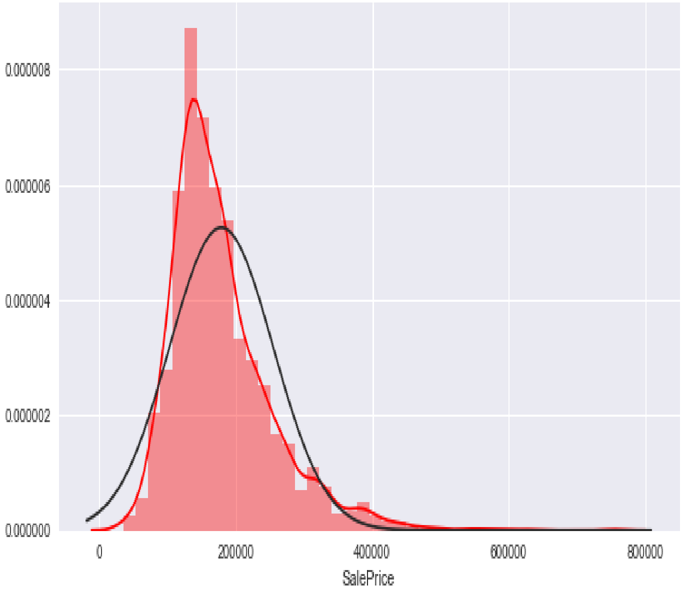
Garage

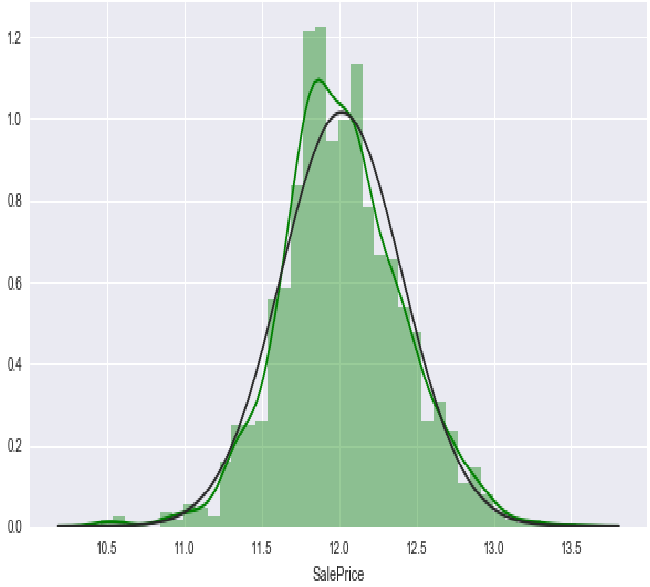
- Most of houses have two cars
- Garage Area makes Chunk having small linearity with Sale Price
- 0 Cars and 1 Cars has no difference in Sale Price
- 4 Cars are similar with 3 Cars house.
- GrLivArea is a good variable not related to Garage Area.

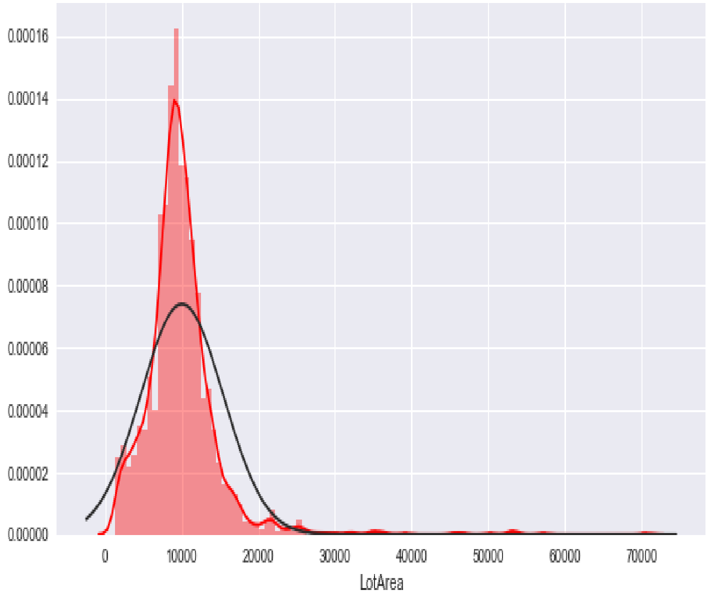


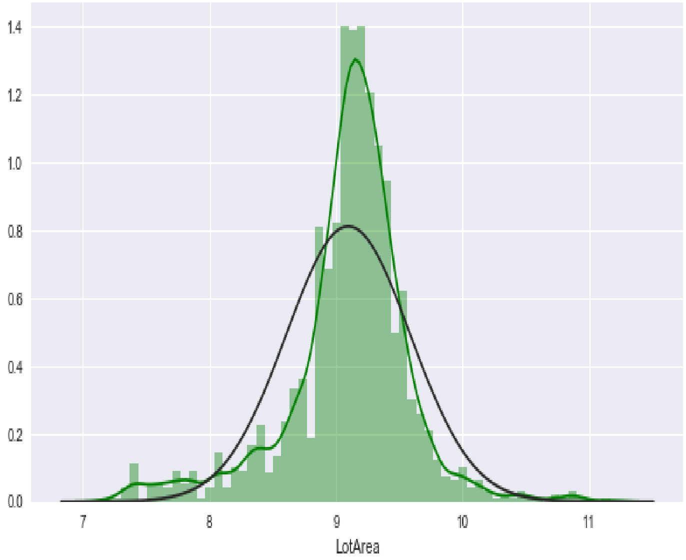
Skewness and Kurtosis

- After applying transformation Skewness of Saleprice:1.6773 was reduced to 0.0608 similarly kurtosis: from 5.2079 to 0.7350
- After applying transformation Skewness of Lot Area : 3.9759 was reduced to -0.7238 similarly kurtosis: from 29.7375 to 2.8932





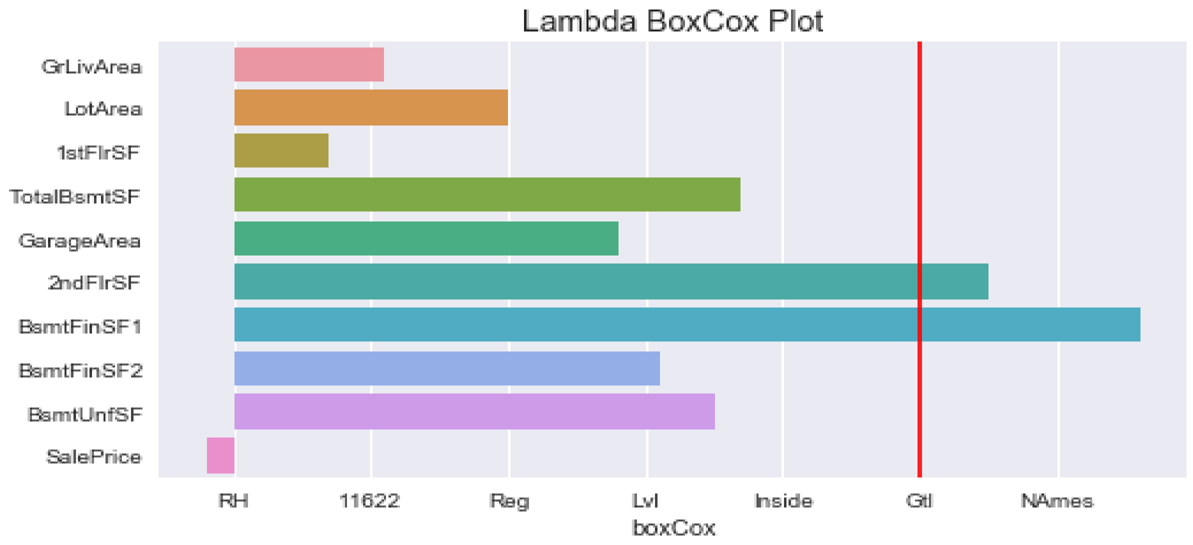


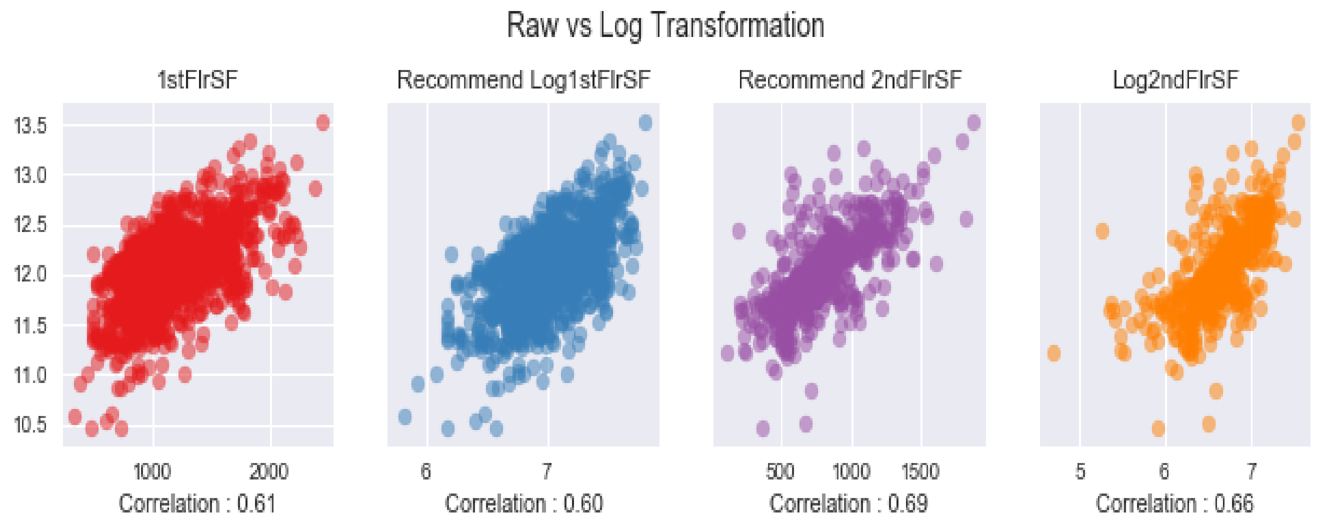


Statistical analysis

Box-cox

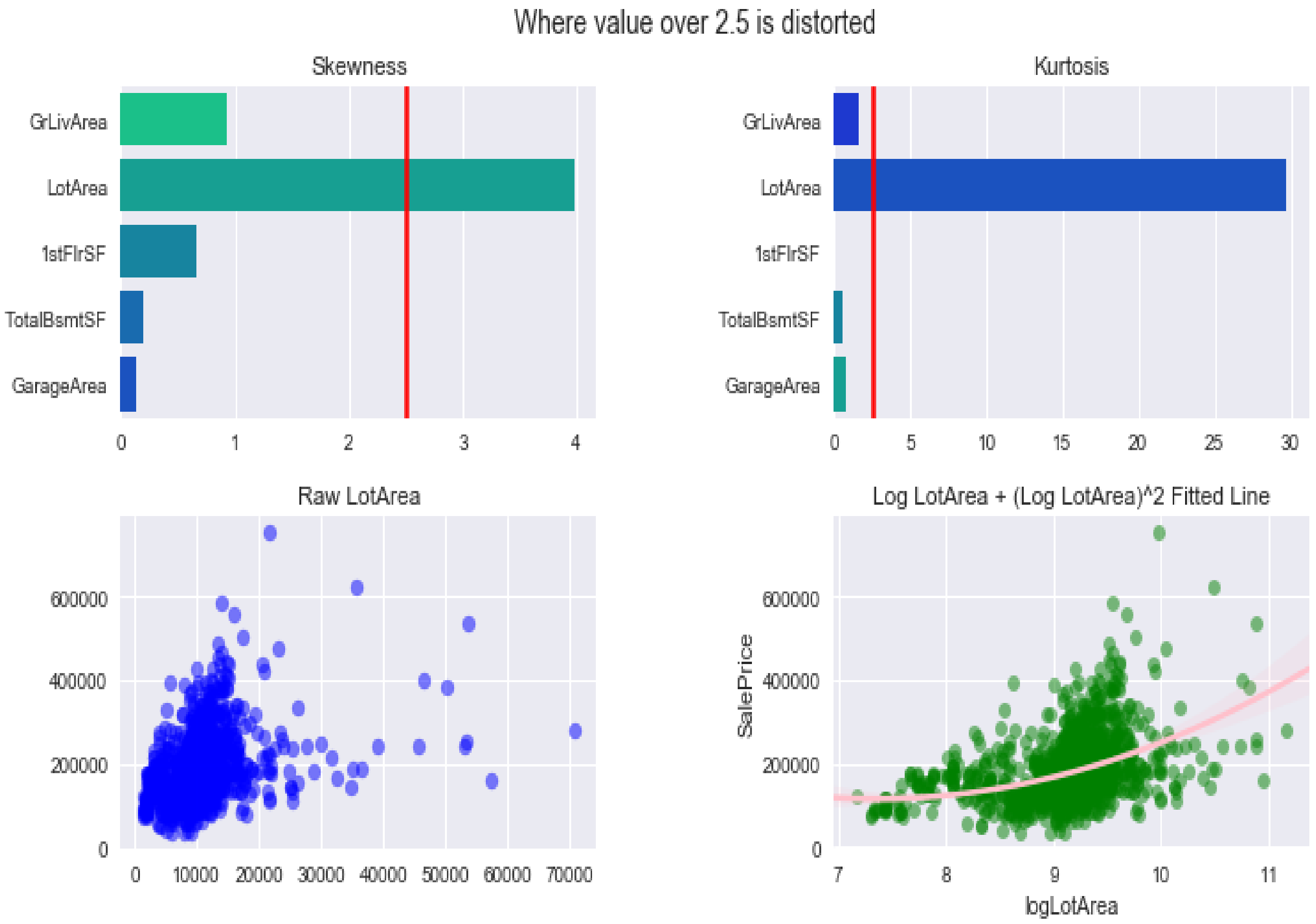
- Except 2ndFlrSF, BsmtFinSF1, the other variables need to deal with by Log Transformation





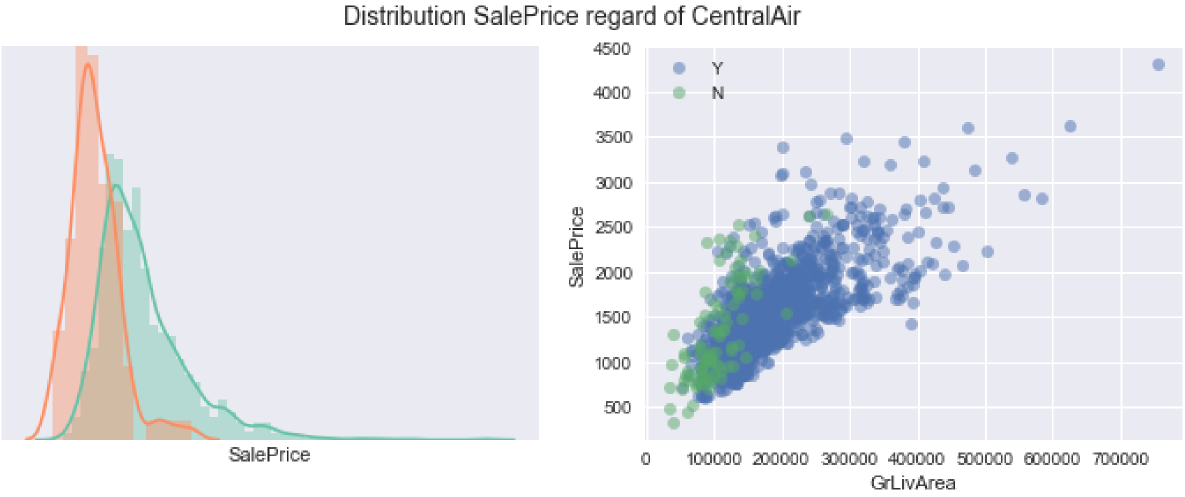
Box-cox

- Here high kurtosis value is most of value gathered in just one part.
- Low value of Lot Area was densely populated
- Just LogLotArea was not good variables, but with (Log LotArea)^2, the fitted line w.r.t Sale Price was better



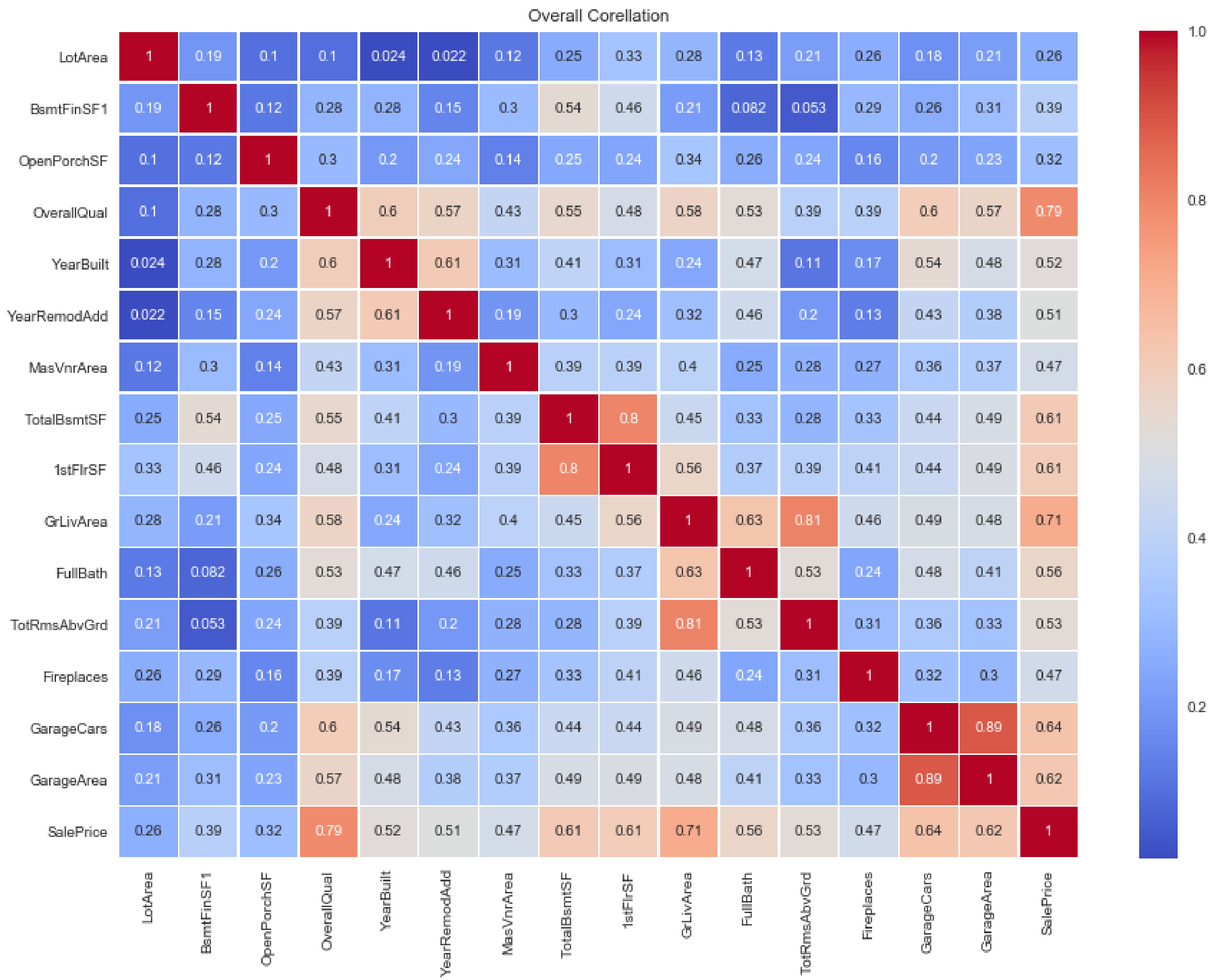
Non-parameteric Test /Wilxoc -rank Sum test

- P_value : 0.00 Sale Price according to Central Air was changed
- Thus Central Air can be used to predict



Correlation

- Year Build and Garage Year Built may just indicate a price inflation over the years.
- There is a strong negative correlation between Basement Unf SF and Basement FinSF2.
- Half Bath and 2nd Floor SF is interesting and may indicate that people gives an importance of not having to rush downstairs in case of urgently having to go to the bathroom
- It can be said that, by essence, some of those features may be combined between each other in order to reduce the number of features (1stFlrSF & Total Bsmt SF, Garage Cars & Garage Area) and others indicates that people expect multiples features to be packaged together.



Feature Engineering

- Feature created based combining many features e.g Total Surface area
- Feature created from existing feature e.g Has Garage
- Feature created from transforming feature i.e Log transformation of Lot Area

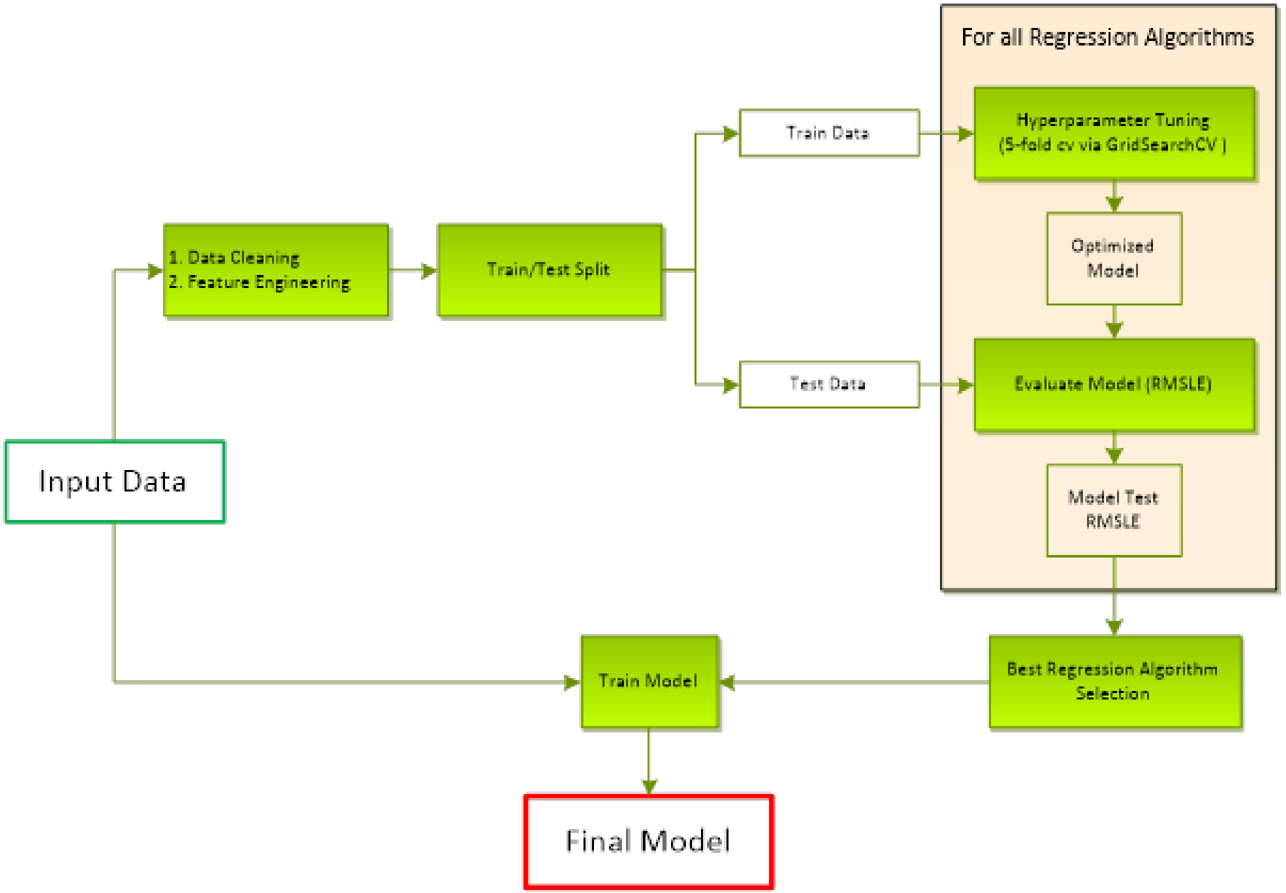
Correlation	TotalBsmtSF	1stFlrSF	2ndFlrSF	'TotalSF'
SalePrice	0.632441	0.613275	0.333395	0.826080

```
full['YearBuiltemodel']=(full['YearBuilt']+full['YearRemodAdd'])
full['TotalSF']=(full['TotalBsmtSF'] + full['1stFlrSF'] + full['2ndFlrSF'])
full['basement']=full['BsmtFinSF2']+full['BsmtUnfSF']
full['Total_sqr_footage'] = (full['BsmtFinSF1'] + full['BsmtFinSF2'] + full['1stFlrSF'] + full['2ndFlrSF'
])

#full['Total_Bathrooms'] = (full['fullBath'] + (0.5 * full['HalfBath']) + full['BsmtfullBath'] + (0.5 * fu
ll['BsmtHalfBath']))
full['Total_Bathrooms'] = (full['FullBath'] + (0.5 * full['HalfBath']) + full['BsmtFullBath'] + (0.5 * ful
l['BsmtHalfBath']))

full['Total_porch_sf'] = (full['OpenPorchSF'] + full['3SsnPorch'] + full['EnclosedPorch'] + full['ScreenPo
rch'] +
                              full['WoodDeckSF'])
full['haspool'] = full['PoolArea'].apply(lambda x: 1 if x > 0 else 0)
full['has2ndfloor'] = full['2ndFlrSF'].apply(lambda x: 1 if x > 0 else 0)
full['hasgarage'] = full['GarageArea'].apply(lambda x: 1 if x > 0 else 0)
full['hasbsmt'] = full['TotalBsmtSF'].apply(lambda x: 1 if x > 0 else 0)
full['hasfireplace'] = full['Fireplaces'].apply(lambda x: 1 if x > 0 else 0)
full["SalePrice"] = np.log1p(full["SalePrice"])
full["LotArea"] = np.log1p(full["LotArea"])
full["BsmtUnfSF"] = np.log1p(full["BsmtUnfSF"])
full["MasVnrArea"] = np.log1p(full["MasVnrArea"])
full["TotalBsmtSF"] = np.log1p(full["TotalBsmtSF"])
full["1stFlrSF"] = np.log1p(full["1stFlrSF"])
full["GrLivArea"] = np.log1p(full["GrLivArea"])
```

Modeling

Modeling Overview

- Type: Supervised Learning
- Pipeline consists of
  - Feature Engineering
  - Train-Test Split
  - Regression Model



LASSO	Value
alpha	0.001
Max_iter	10000

Ridge	Value
alpha	0.3
Max_iter	100000

Random Forest	Value
n_estimators	10000
max_depth	6
max_features	None
min_samples_leaf	3
min_samples_split	12

Scikit Learn Model Design

LightGM	Value
objective	regression
num_leaves	5
learning_rate	0.01
n_estimators	4000
max_bin	200
bagging_fraction	0.7
bagging_freq	5
feature_fraction	0.1
verbose	-1

Gradient Boost	Value
n_estimators	3000
max_depth	4
max_features	sqrt
min_samples_leaf	15
min_samples_split	15
learning_rate	0.01
loss	huber

Scikit Learn Model Design

XgBoost	Value
objective	reg:linear
n_estimators	2500
learning_rate	0.015
max_depth	3
min_child_weight	0
gamma	0
subsample	0.6
colsample_bytree	0.6
scale_pos_weight	1

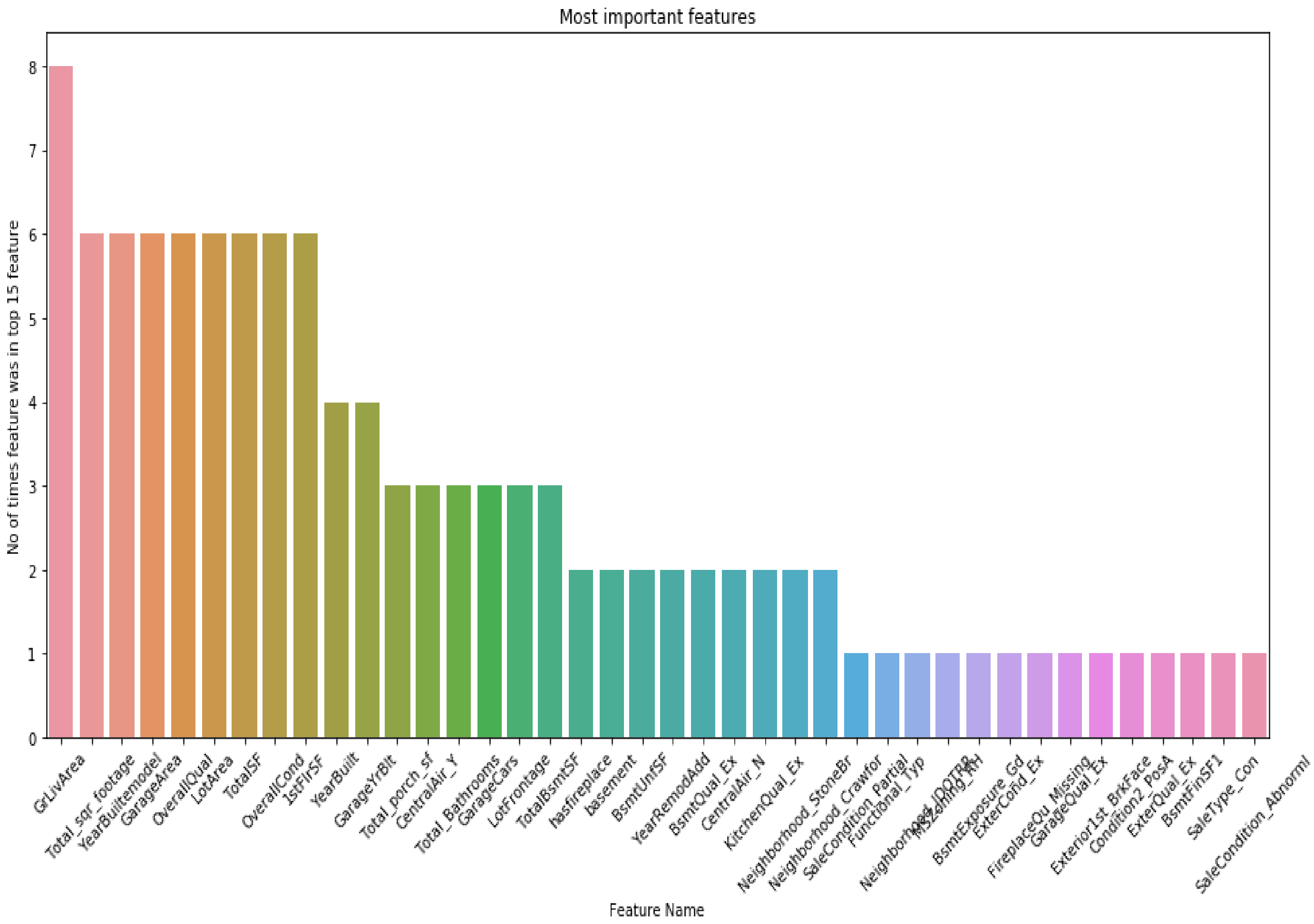
ElasticNet	Value
alpha	0.0007
Max_iter	30000

ADA Boost	Value
loss	linear
n_estimators	4000
learning_rate	1.1

Scikit Learn Model Design

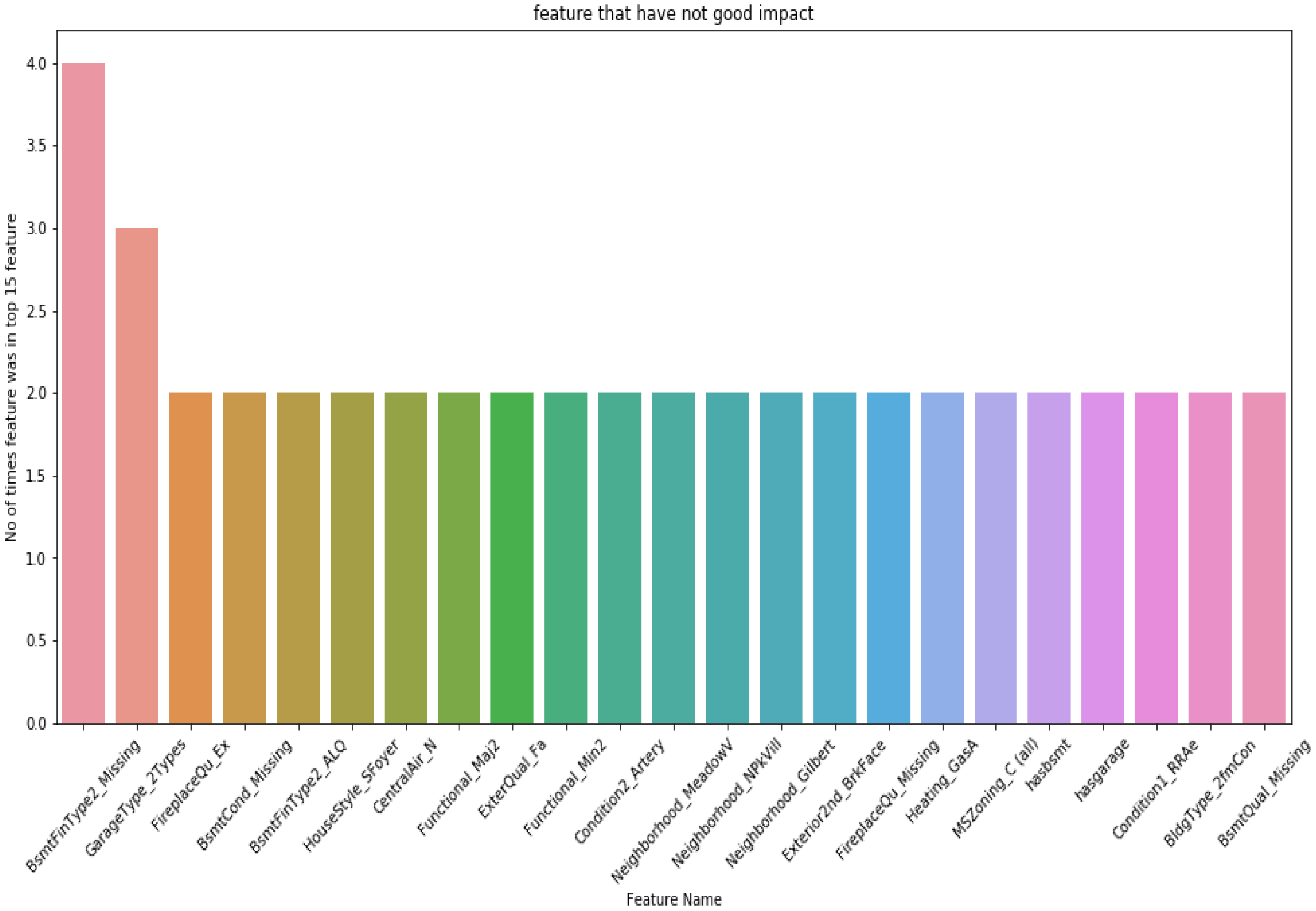
Most important Feature

- most important factors that affect positive impact on house price are Area, Quality, how old is house
- living area, garage, porch, air conditioning and bathroom also have big impact
- fireplace , basement , having nice neighborhood and sale type or condition is also a good predictor
- Feature engineered features like Total_Sql_footage, Total SF are very good parameter
- parameter like has fireplace is more important then area or quality of fireplace



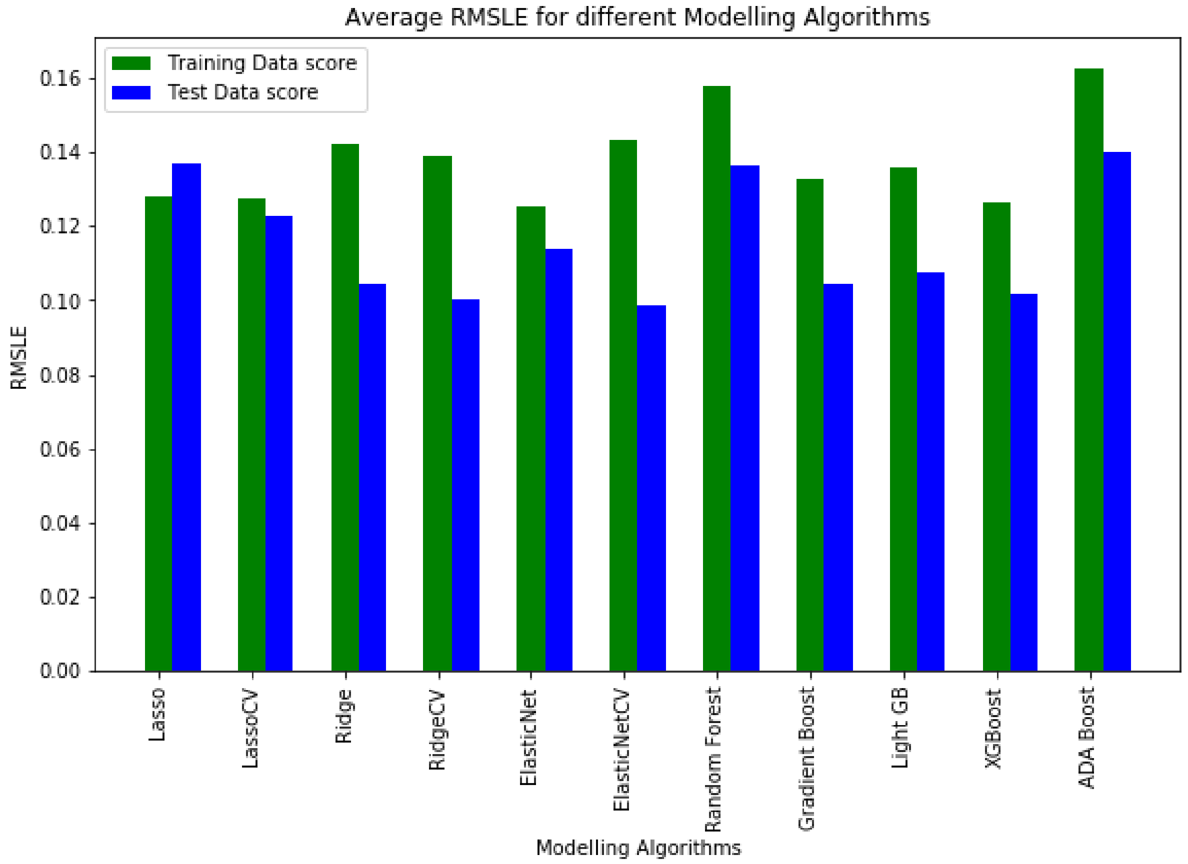
Least important Feature

- house without basement and house with two different type of garage has negligible impact on house price
- Even though having fireplace can be good, people do ignore what is quality of fireplace
- not having garage, basement or any other section does not have any positive impact



Model Summary

- Best algorithms which can be used are Lasso, XgBoost, Elastic Net
- Random forest and Ada Boost are not very good predictors
- execution time of three selected algorithm is also less
- ADA boost and Random Forest are not very good predictors or may need more hyper parameter tuning
- Gradient Boost and Light Gradient boost are not bed predictors either



Thank you